°textA dissertation submitted in partial fulfillment of the requirements for the degree of

Certain observations concerning the effects of epistasis on complex traits and the evolution of genomes.

Maximilian Press

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Christine Queitsch, Chair

Elhanan Borenstein, Chair

First committee member

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Certain observations concerning the effects of epistasis on complex traits and the evolution of genomes.

Maximilian Press

Co-Chairs of the Supervisory Committee:
Associate Professor Christine Queitsch
Department of Genome Sciences

Associate Professor Elhanan Borenstein
Department of Genome Sciences

The informational content of genomes is usually interpreted mimetically, which is to say, with a one-to-one relationship between genotypes at certain genomic positions and phenotypic outcomes. While such interpretations have the virtue of simplicity, they are empirically unsuccessful in elucidating the working of biological systems and in allowing us to predict variation in many phenotypes. Many have called for such models to explicitly consider epistasis, which can roughly be defined as any consideration of interactions between genomic elements (a semiotic rather than mimetic interpretation of genomic information). In this thesis, I consider some of the consequences of the existence of such epistasis. In the first part of this thesis, I consider a particular case of a fast-evolving genetic element (short tandem repeat, or microsatellite) that shows widespread epistasis, and propose that such elements are likely to accumulate epistatic interactions by acting as mutational modifiers. I go on to show molecular mechanisms by which the element participates in epistasis, their phenotypic consequences, and make some observations on other short tandem repeats. In the second part of this thesis, I start with the assumption of epistasis between genes, and explore how this assumption can be used to understand the evolution of bacterial genome con-

tent. First, I take Hsp90, the known epistatic hub, and infer its coevolution with other genes through coordinated gains and losses across bacterial diversity. I further extend the underlying phylogenetic model to predict new "clients" of bacterial Hsp90, which have remained elusive when pursued through purely experimental approaches. Collaborators were able to validate certain of these predicted clients. Last, I attempt an analogy between prokaryotic genome evolution and the much better-understood field of protein evolution. I propose that, like protein evolution by substitution, genome evolution by horizontal acquisition of genes is substantially constrained by epistasis. I go on to infer the existence of such epistatic dependencies, where one gene in an ancestral genome promotes the acquisition of a second gene. A network of such dependencies shows a chronological structuring of gene acquisitions through prokaryotic evolution, suggesting universal assembly patterns by which genomes acquire functions. I go on to show that these dependencies are taxonomically universal (i.e. not restricted to particular phyla), and that they are sufficient to make reasonably good predictions about what genes a genome will gain in the future. This predictability of genome evolution by horizontal transfer confirms a major assertion of the protein evolutionists, that constraining epistasis leads to predictable evolutionary outcomes. Together, these observations indicate that the genetic architecture of traits and the content of genomes are shaped by the existence of networks of epistasis, reflecting the complex wiring of underlying biological functions.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

"I will ask you to mark again that rather typical feature of the development of our subject; how so much progress depends on the interplay of techniques, discoveries and new ideas, probably in that order of decreasing importance."

Sydney Brenner

That generation's dream, aviled
In the mud, in Monday's dirty light,

That's it, the only dream they knew,
Time in its final block, not time

To come, a wrangling of two dreams.
Here is the bread of time to come,

Here is its actual stone. The bread
Will be our bread, the stone will be

Our bed and we shall sleep by night.
We shall forget by day, except

The moments when we choose to play
the imagined pine, the imagined jay.

Wallace Stevens

DEDICATION

to my Sarah

Chapter 1

INTRODUCTION

1.1 Practical and philosophical approaches to the genotype-phenotype problem

[8]

1.2 Linear models and Fisherian quantitative genetics

1.3 Epistasis

1.4 Fast-evolving genetic elements

1.4.1 Short tandem repeats (STRs)

1.4.2 Horizontal acquisition of genetic material

1.5 Evolutionary implications of fast-evolving genome architectures

1.5.1 'Tinkering' and duct tape in evolution

1.5.2 Parallelism and Predictability

Chapter 2

BACKGROUND-DEPENDENT EFFECTS OF POLYGLUTAMINE VARIATION IN THE *ARABIDOPSIS THALIANA* GENE *ELF3*

A version of this chapter was published under the following reference:

Soledad F. Undurraga, Maximilian O. Press, Matthieu Legendre, Nora Bujdoso, Jacob Bale, Hui Wang, Seth J. Davis, Kevin J. Verstrepen, and Christine Queitsch. Background-dependent effects of polyglutamine variation in the *Arabidopsis thaliana* gene *ELF3*. Proceedings of the National Academy of Sciences of the United States of America, 109(47):19363-7, November 2012.

Soledad Undurraga, Jacob Bale, Nora Bujdoso, and Hui Wang contributed transgenic lines, experiments, and figures.

Supplementary figures and tables can be found in Appendix A.

2.1 Summary

Tandem repeats (TRs) have extremely high mutation rates and are often considered to be neutrally evolving DNA. However, in coding regions, TR copy number mutations can significantly affect phenotype and may facilitate rapid adaptation to new environments. In several human genes, TR copy number mutations that expand polyglutamine (polyQ) tracts beyond a certain threshold cause incurable neurodegenerative diseases. PolyQ-containing proteins exist at a considerable frequency in eukaryotes, yet the phenotypic consequences of natural variation in polyQ tracts that are not associated with disease remain largely unknown. Here, we use Arabidopsis thaliana to dissect the phenotypic consequences of natural variation in the polyQ tract encoded

by *EARLY FLOWERING 3* (*ELF3*), a key developmental gene. Changing ELF3 polyQ tract length affected complex ELF3-dependent phenotypes in a striking and non-linear manner. Some natural ELF3 polyQ variants phenocopied elf3-loss-function mutants in a common reference background, although they are functional in their native genetic backgrounds. To test the existence of background-specific modifiers, we compared the phenotypic effects of ELF3 polyQ variants between two divergent backgrounds, Col and Ws, and found dramatic differences. Our data support a model in which variable polyQ tracts drive adaptation to internal genetic environments.

2.2 *Introduction*

In coding regions, tandem repeat (TR) copy number variation can have profound phenotypic effects [3]. For example, TR copy number mutations that expand polyglutamine (polyQ) tracts past a threshold number of glutamines can cause incurable neurodegenerative diseases such as Huntington's disease and Spinocerebellar Ataxias [2, 7]. PolyQ tract length correlates with onset and severity of polyQ expansion disorders, but for intermediate polyQ tracts this correlation is far weaker (4-8), suggesting the existence of genetic and environmental modifiers (9-12). Despite their potential for pathogenicity, variable polyQ tracts occur frequently in eukaryotic proteins, many of them functioning in development and transcription (1, 13-15). Model organism studies have suggested that coding TRs are an important source of quantitative genetic variation that facilitates evolutionary adaptation (1, 16-19). For example, TR copy number variation in the yeast gene FLO1 correlates linearly with flocculation (20), a phenotype that is important for stress survival (17). As polyQ tracts often mediate protein interactions (2, 3, 21), polyQ-encoding TR copy number mutations could produce large and possibly adaptive phenotypic shifts. To determine the phenotypic impact of naturally occurring polyQ variation (18, 22, 23) in a genetically tractable model, we focused on the gene ELF3, which encodes a polyQ tract that is highly variable across divergent Arabidopsis thaliana strains (accessions) (19, 24). ELF3 is a core component of the cir-

cadian clock and a potent repressor of flowering, and is considered a "hub protein" for its many interactions with various proteins (24-31). Consequently, elf3 loss-of-function mutants show pleiotropic phenotypes: they flower early, show poor circadian function, and grow long embryonic stems (hypocotyls) in light (25-27, 29, 30, 32). Single nucleotide polymorphisms (SNPs) in ELF3 affect shade avoidance, a fitness-relevant plant trait (24, 33). ELF3 polyQ variation has been suggested to correlate with two parameters of the circadian clock, period and phase (19). The ELF3 polyQ tract may mediate ELF3 membership in protein complexes, though thus far no ELF3-binding protein is known to bind it (26, 28-30). We discovered that altering polyQ tract length has dramatic effects on ELF3-dependent phenotypes and that these effects are dependent on genetic background.

2.3   Methods

2.3.1   Plant Materials and Growth Conditions.

The 181 Arabidopsis thaliana accessions are as previously described (1). The loss-of-function EARLY FLOWERING 3 (elf3) mutants are: (i) *elf3-4*, containing a CCR2::LUC transgene (ecotype Ws) (2, 3) and 2) *elf3-200*, the GABI750E02 T-DNA insertion mutant (ecotype Col-0) (4). For hypocotyl experiments, seeds were sterilized with Ethanol and plated onto 1x Murashige and Skoog (MS) basal salt medium supplemented with 1x MS vitamins, 1% sucrose, 0.05% Mes (wt/vol), and 0.24% (wt/vol) phytagel. After stratification in the dark at 4° C for 3 d, plates were transferred to an incubator (Conviron) that was set to either short day (SD) (8L:16D at 20° C) or long day (LD) (16L:8D at 22° C : 20° C), with light supplied at 100μmol $* $ m² $* $ s^{-1} by cool-white fluorescent bulbs. For growth on soil, seeds were stratified at 4° C for 3 d, and then grown in Sunshine #4 soil under cool-white fluorescent light at either LD or SD at 20 °C. Seedlings used for RNA extractions were grown on soil under LD conditions and harvested on day 10. Samples for ELF3 expression measurements were collected at Zeitgeber time (ZT) 20. Samples

for Phytochrome- interacting Factor 5 (PIF5) expression measurements were collected a ZT 8. Samples for and Pseudoresponse regulator 9 (PRR9) expression measurements were collected at ZT 0, 5, and 8.

2.3.2 Generation of ELF3 Transgenic Plants.

To generate A. thaliana transgenics carrying different ELF3 tandem repeat (TR) alleles, the cDNA clone RAFL09-28-E05 (RIKEN BRC) (5, 6), containing the ELF3 coding region and 3' UTR (Col-0 accession) was used. This cDNA clone lacks the small 5' intron. Two re- striction sites, NarI and NcoI, were inserted into the ELF3 coding sequence using the QuikChange Site-Directed Muta- genesis kit (Stratagene) (primer information in Table A4). The polyglutamine (polyQ)-encoding region was amplified from accessions containing selected TR copy number alleles (primer information in Table A4, TR allele information in Table A1). These PCR products were digested with NarI/NcoI and ligated into the previously mutagenized ELF3 coding region. An artifi- cial allele lacking the TR was generated by site-directed muta- genesis (primer information in Table A4). Mutated plasmids and all ligation products were sequenced to ensure accuracy. The ELF3 alleles were cloned into pENTR1A (Invitrogen). A 2-kbp NotI fragment containing the ELF3 promoter was inserted up- stream of each ELF3 coding sequence. The fragments containing the ELF3 promoter, ELF3 coding sequence, and the ELF3 3' UTR were recombined using Gateway LR Clonase II (In- vitrogen) into a modified pB7WG2 (7), which lacks the CaMV- 35S promoter. The region encoding the polyQ tract of each construct was sequenced to ensure accurate TR copy number. The plasmids were used to transform Agrobacterium tumefaciens GV3101. Subsequently, Arabidopsis elf3 mutants were trans- formed by the flower dip method (8). Transformants were se- lected on Basta (Liberty herbicide; Bayer Crop Science) and propagated for three to four generations. The accuracy of the transgenes was confirmed by PCR (primer information in Table A4). All Ws phenotypic assays were performed in homozygous transgenic plants with expression levels between 0.8- and 4.5- times the respective ELF3 wild-type (Fig-

ure A1C); for Col lines, transgene expression levels were between 0.3- and 4.3-times the respective ELF3 wild-type (Figure A1D). Analyzed plant lines are in Tables A2-A4.

2.3.3 RNA Extractions and Real-Time PCR.

Total RNA was extracted from 30-mg frozen tissue using the SV Total RNA Isolation System (Promega). Subsequently, 2 μg of RNA were subjected to DNase treatment using Ambion Turbo DNA-free Kit (Applied Biosystems). RNA integrity and purity were checked with an Agilent Bioanalyzer using the RNA 6000 Nano Kit (Agilent Technologies). For cDNA synthesis, 200 ng of DNase-treated RNA was reverse-transcribed using the Transcriptor First Strand cDNA Synthesis Kit (Roche) and oligo dT primers. Transcript abundance was determined by real-time quantitative PCR using the Light-Cycler 480 system (Roche), with LightCycler 480 SYBR Green I Master (Roche) and the following PCR conditions: 5 min at 95 °C, followed by 35 cycles of 15 s at 95 °C, 20 s at 55 °C, and 20 s at 72 °C. To ensure that PCR products were unique, a melting- curve analysis was performed after the amplification. UBC21 expression (At5g25760) was used as a reference. All quantitative RT- PCR primers were designed with the LightCycler Probe Design Software (Roche). Sequences for real-time PCR primers are shown in Table A4. Relative quantification was determined with the $\Delta\Delta C_T$ Method (9). Error was calculated as previously described (10).

2.3.4 Thermal Asymmetric Interlaced PCR.

High-efficiency thermal asymmetric interlaced (TAIL)-PCR was performed as previously described (11) to obtain the flanking sequence of the construct in- tegration site (left border). Briefly, a preamplification step was performed with primers LAD and LB-0a (Table A4), followed by primary TAIL-PCR with primers AC1 (11) and LB-1a (Table A4), and 1 μL of a 1/40 dilution of the preamplification product as a template. A secondary TAIL-PCR with primers AC2 (11) and LB-2a (Table A3) was performed with 1 μL of a 1/10 dilution of the primary TAIL-PCR product. Next, 3-kbp products

were extracted from agarose gels and subsequently Sanger-sequenced. Only sequences containing the T-DNA left border were considered.

2.3.5 *Developmental Phenotype Assays.*

For measurements of hypocotyl length, seedlings were grown on vertical plates for 15 d in a pseudorandomized design under either SD or LD conditions (12). Hypocotyl length was measured with ImageJ on digital images (http://rsbweb.nih.gov/ij/). For measurement of flowering time, seeds were planted in sheet pots (36 pots per tray) in a randomized design and trays were rotated daily. Flowering time was re- corded as the day when the inflorescence reached 1 cm in height. Rosette leaf number was determined on the same day. Petiole- length/leaf-length (PL/LL) ratio for leaf four was determined on day 45. Least-square means for all traits were derived from a linear regression analysis for each trait separately. ELF3 TR copy number was modeled as a nominal variable and independent transgenic lines carrying the same ELF3 TR allele were analyzed together. We tested for significant phenotypic differences con- ferred by the different ELF3 TR alleles by using Tukey-HSD tests with $\alpha = 0.05$ that accommodate nonnormal data.

2.3.6 *Luciferase Imaging and Period Analysis.*

Luciferase assays were performed with lines containing the CCR2::LUC reporter. Seeds were surface sterilized with a 70% (vol/vol) ethanol wash followed by a second wash with 33% (vol/vol) Klorix with Triton X-100, and then rinsed twice with sterile water. Seeds were plated on MS3 medium [pH 5.7, 3% (wt/vol) sucrose, 1.5% (wt/vol) PhytoAgar, and 15 μg/mL hygromycin B]. They were subsequently stratified for 4 d at 4 °C in the dark and entrained under 12-h light:12-h dark cycles under white fluorescent light ($\sim$ 10μmol $* $m^{-2} $*$ s^1) at 22 °C. On the sixth day, a minimum of 24 seedlings per line was transferred to 96-well TopCount (Perkin-Elmer) plates containing 200 mg MS3 agar. We added 5 mM Luciferin in 0.01% Triton X-100 and entrained seedlings for another cycle before luminescence was detected using a Packard/Perkin-Elmer Top-

Count Scintillation and Luminescence Counter. Red and blue light-emitting diodes (100μmol $* $ m^{-2} $* $ s^{-1}) were used as a light source during this analysis. During the first 24 h of luminescence detection, plants were grown in 12-h light:12-h dark and then released under constant light conditions to measure the free- running period. Each individual was measured approximately every 30 min for a minimum of 5 d. Luminescence levels were quantified and analyzed as previously described (2, 3) using the macro suites TopTempII and Biological Rhythms Analysis Software System (13). Period length and relative amplitude error (RAE) were estimated using fast Fourier transform nonlinear least squares (14). Period values scored with RAE values below 0.4 were considered robustly rhythmic (15).

2.3.7   *Principal Component Analysis.*

We clustered our phenotypic data using principal component analysis (PCA) to find patterns corresponding to genotypes. We excluded the phenotype of rosette leaf number in SD, for which data were missing for several alleles. The phenotypes included in the analysis are: Days to flowering in SD and LD conditions, hypocotyl length under SD and LD PL/LL for the fourth leaf in SD, and rosette leaf number in LD. For analyses involving Col lines, the SD PL/LL ratio phenotype was omitted because of lack of data, and PCA was thus based on the remaining five phenotypic variables. For each phenotype in each genetic background (either Ws or Col-0), we calculated the mean phenotype of the independently generated lines for each *ELF3-TR* allele, giving us a 28 x 6 matrix of mean phenotypes for the 28 genotypes for each of six phenotypic variables. Within each background, we ranked the genotypes for each phenotype. Ranks were transformed into a standard normal distribution based on their percentile, using the R function qnorm. Using this transformed dataset, we performed PCA using the R function *prcomp* (R Foundation for Statistical Computing, http://www.r-project.org/, 2011). We performed PCA for each background separately, and then for both backgrounds together. Rank-normalization was necessary to compare (i) phenotypes measured on

different scales and (ii) Ws- and Col-derived plants, between which backgrounds absolute phenotypic differences exist. Consequently, the rank-normalization increases stability of our estimates, as our dataset is relatively small and PCA□ as assumptions of normality were not met by our raw dataset. PCA on raw values scaled to a standard normal distribution gave similar results. Biplots were generated with the R *biplot* function on *prcomp* function output.

2.4 ELF3-TR variation affects ELF3-dependent phenotypes.

Among 181 natural A. thaliana accessions, the *ELF3-TR* encoded between 7 and 29Q (Table A1, Figure A1a). For comparison, polyQ expansions over 20Q are associated with disease in the context of the SCA6 gene, though most other disease-associated polyQ expansions are longer (2, 19, 24). The most frequent *ELF3-TR* encoded 16Q, whereas the shortest TR (7Q) was found in the reference strain Col-0. We set out to test whether naturally occurring *ELF3-TR* alleles affect ELF3-dependent phenotypes and whether they do so in a linear manner as suggested by association studies (19) and found for coding TR variation in other genes (16, 20). We generated expression-matched transgenic lines for most natural *ELF3-TR* alleles in the loss-of-function *elf3-4* mutant (Ws background, Table A2, Figure A1c) (32) and measured their flowering time and circadian clock-related phenotypes (Figures 1, SAa-g). *ELF3-TR* variation significantly affected ELF3-dependent phenotypes, but there was no evidence of a linear relationship. The different *ELF3-TR* alleles resulted in phenotypes ranging from nearly full complementation of *elf3-4* to nearly phenocopying the loss-of-function mutant. We used principal components analysis to describe the complex effects of *ELF3-TR* alleles on all tested ELF3-dependent phenotypes (PCA, Figures 1a, A2h-j). Principal component 1 (PC1) corresponds to general functionality of ELF3 in all measured phenotypes, with wild-type Ws and mutant *elf3-4* defining the extremes. Separation along PC1 is driven by the tendency of plants with functional ELF3 to show short hypocotyls, late flowering, increased rosette leaf number, and short petioles (Figures 1b-d, A2). The en-

dogenous ELF3-16Q allele complemented both the early-flowering and long-hypocotyl phenotypes of *elf3-4* (Figures 1b-d, A2). In contrast, both the long ELF3-23Q and the short ELF3-7Q allele (endogenous TR alleles in Br-0/Bur-0 and Col-0, respectively) behaved similarly to the *elf3-4* loss-of-function allele (Figures 1b-d, A2), although they are functional in their native backgrounds. Neither Col-0 nor Br-0 and Bur-0 show the phenotypic characteristics of elf3-mutants (early flowering (34), long hypocotyls (35) and long petioles (36)), suggesting that *ELF3-TR* alleles may interact with background-specific modifiers. ELF3-0Q, an artificial ELF3 allele lacking the TR, partially complemented *elf3-4* (Figures 1a, S2). Hence, the polyQ-encoding TR is not necessary for all ELF3 function, but changes in TR copy number are sufficient to enhance or ablate ELF3 function.
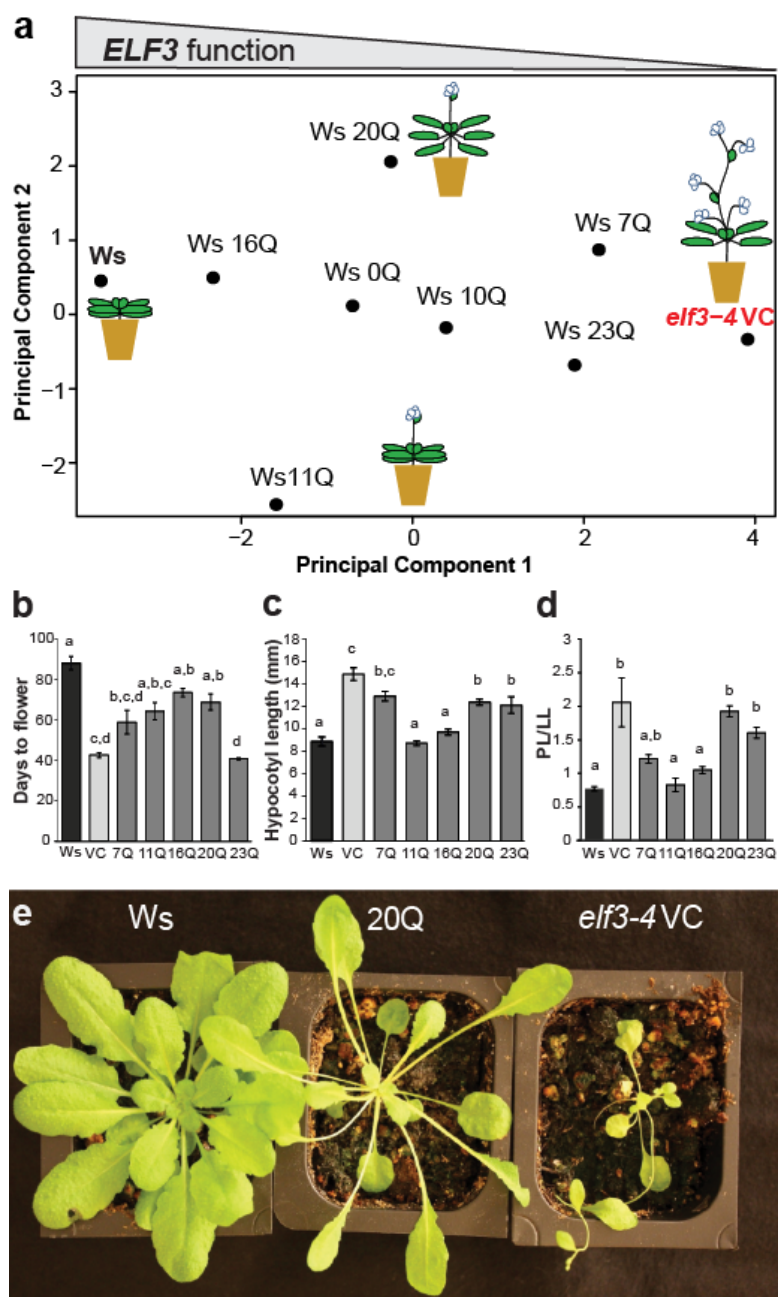
Figure 2.1

Figure 2.1: ELF3-TR variation has nonlinear phenotypic effects. (A) PCA of developmental traits of all ELF3-TR copy number variants. A. thaliana images illustrate ELF3-TR effects on the traits days to flower and hypocotyl length under SD and LD, petiole-length/leaf-length ratio (PL/LL) under SD only, and rosette leaf number under LD only. The contributions of specific phenotypes to PCs are in Figure A2J. Representative TR copy number alleles are shown from an analysis including all alleles (for all alleles see Figure A2 H and I). (B) Days to flower under SD conditions for selected lines. n = 6 plants per transgenic line. (C) Hypocotyl length at 15 d under SD for selected lines. n = 20-30 seedlings per transgenic line. (D) PL/LL of the fourth leaf for selected lines. Data are from the same plants as in B. (E) Plants carrying the ELF3-20Q allele (Center) are specific hypomorphs under SD with the elongated petioles of the elf3-4 mutant (vector control, VC, Right) and a wild-type flowering phenotype (Ws, Left). ELF3-TR alleles are indicated with the number of Qs encoded, Ws is wild-type, VC is the elf3-4 vector control. Error bars are SEMs. Genotypes labeled with different letters differed significantly in phenotype by Tukey's HSD test. For all Ws-background phenotype data, see Figure A2 A□G. Data are from multiple independently generated expression-matched (Figure A1C) T3 and T4 lines for each TR copy number allele (Tables A2, A3). These experiments were repeated at least once with similar results. The tested ELF3-20Q lines contained unique insertions that did not affect genes with known function.

PC2 separated ELF3-20Q and ELF3-11Q, which behaved as hypomorphs in certain phenotypes but not others (Figure 1a). For example, ELF3-20Q plants had significantly longer hypocotyls than wild-type and its petioles phenocopied the extremely long petioles of the *elf3-4* mutant (Figure 1c-e), but they did not differ from wild-type in flowering time (days to flower, Figure 1b). The existence of both general and specific hypomorphs suggests that polyQ variation affects the multiple ELF3 functions separately. As part of a protein complex, ELF3 affects expression of Phytochrome-interacting Factor 5 (PIF5) and Pseudo-response regulator 9 (PRR9) (28, 37, 38). PIF5 and PRR9 expression were

strongly affected by ELF3 polyQ variation (Figure A3). ELF3-16Q phenocopied wild-type PRR9 and PIF5 expression, and the hypomorphic ELF3-23Q phenocopied *elf3-4* (28, 37, 38), mirroring their developmental phenotypes. Consistent with their divergence along PC2 (Figure 1a), ELF3-11Q and ELF3-20Q differed in their effect on PRR9 expression, but not on PIF5 expression (Figure A3a, b), demonstrating that ELF3 polyQ variation differentially affects the regulation of downstream genes.

2.5   ELF3-TR variation modulates the precision of the circadian clock

To directly assess the role of ELF3 polyQ variation in the circadian clock, we used the CCR2::LUC reporter system (25, 39). We observed little difference in circadian period among wild-type Ws and tested ELF3-TR alleles (Figure A4a), contradicting a previously observed association of TR copy number with period in natural accessions (19). However, we found that the relative amplitude error (RAE) of oscillation varies substantially across ELF3-TR genotypes (Figures 2a, S4b). RAE measures the precision of a circadian period (40): high RAE values (> 0.4) indicate poor oscillation and clock dysfunction (41). The endogenous Ws ELF3-16Q nearly complemented the elf3-4 RAE defect, whereas the TR alleles ELF3-7Q, ELF3-10Q, and ELF3-23Q showed higher RAE, approaching arrhythmic *elf3-4* levels (Figure 2a, b), consistent with their hypomorphic performance in other ELF3 traits (close to *elf3-4* in PC1, Figure 1a). Together, these results suggest that ELF3 polyQ tract length is a critical determinant of circadian clock precision, but not period length, in A. thaliana.

2.6   ELF3-TR variation interacts with genetic background.

To test our hypothesis that *ELF3-TR* variation interacts with genetic background, we regenerated all *ELF3-TR* transgenic lines in the *elf3-200* loss-of-function mutant with matched transgene expression (Col background, Table A3, Figure A1d) (42). We used PCA to compare *ELF3-TR* effects between Ws and Col backgrounds (Figures 3a, S5). The Col-specific ELF3-7Q allele complemented *elf3-200* in some traits such as flower-
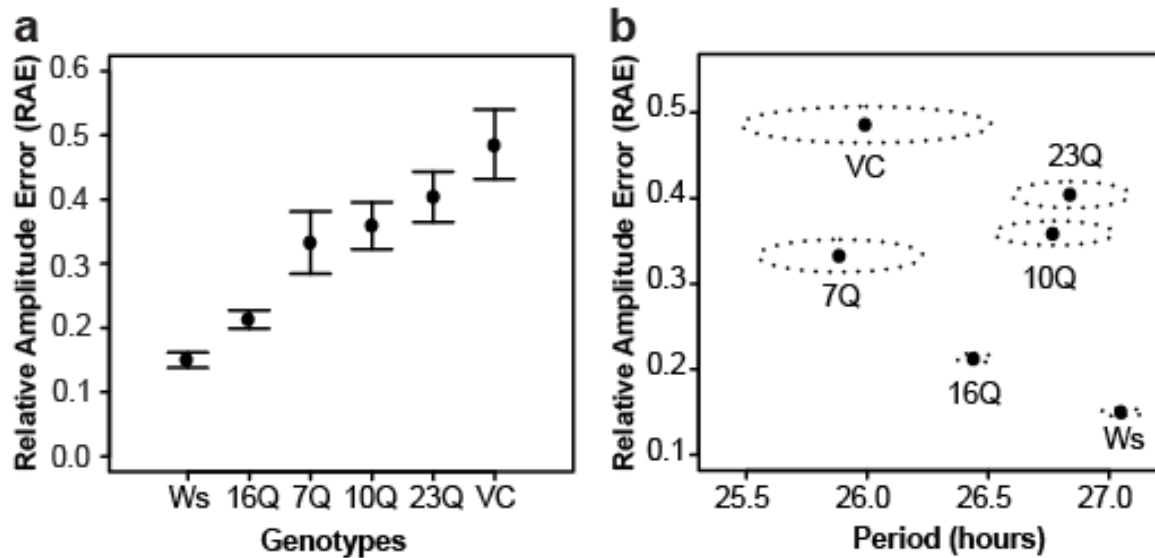
Figure 2.2: ELF3-TR variation modulates the precision of the circadian clock. (A) RAE of CCR2::LUC circadian oscillation in seedlings with indicated ELF3-TR alleles. Bars represent 99% confidence intervals. (B) Mean values of circadian period and RAE (points) were measured in seedlings with indicated ELF3-TR alleles. Dotted ellipses represent SEMs for both period and RAE. Note that plants with high RAE have extremely unreliable estimates of circadian period. Bioluminescence rhythms from the CCR2::LUC reporter in ELF3-TR transgenic lines were used to measure circadian parameters under LL after 5 d of entrainment in 12-h light:12-h dark cycles. n≥100 seedlings for all genotypes. Aggregate data from four independent experiments are shown. See Fig. A4 for RAE and period data for all alleles.

ing time (in short days, SD) and hypocotyl length (in long days, LD), but not others (Figures 3a, b, S5, S6). This result may be due to the absence of the small 5' intron from the ELF3 construct used in this study. However, there was still a dramatic spread of phenotypes: all longer *ELF3-TR* alleles (>20 Qs) nearly complemented *elf3-200*, delaying flowering and shortening hypocotyls, whereas few of the shorter alleles did (Figures 3, S5, S6). Results were similar when the Col data were analyzed alone (Figure A6). Thus, in contrast to our results in the Ws background, *ELF3-TR*s appeared to show a threshold effect for TR copy number in the Col background. We speculate that the intensive laboratory propagation of the Col-0 accession may have altered selection on the *ELF3-TR*, resulting in an extremely short "hypomorphic" allele, whereas under natural conditions a longer TR might be more functional. Comparing TR allele effects between the two backgrounds revealed striking differences. For example, the ELF3-23Q allele was generally hypomorphic in the Ws background (*elf3-4*), whereas it produced highly functional ELF3 in the Col background (*elf3-200*, Figure 3). In turn, the ELF3-16Q allele produced highly functional ELF3 in the Ws background (*elf3-4*), but was generally hypomorphic in the Col background (*elf3-200*). The consistent performance of the artificial ELF3-0Q allele across backgrounds suggests that the background effect is TR-dependent (Figures 3a, S5). Collectively, our results support that *ELF3-TR* alleles interact with background-specific modifiers.

a

ELF3 function

b
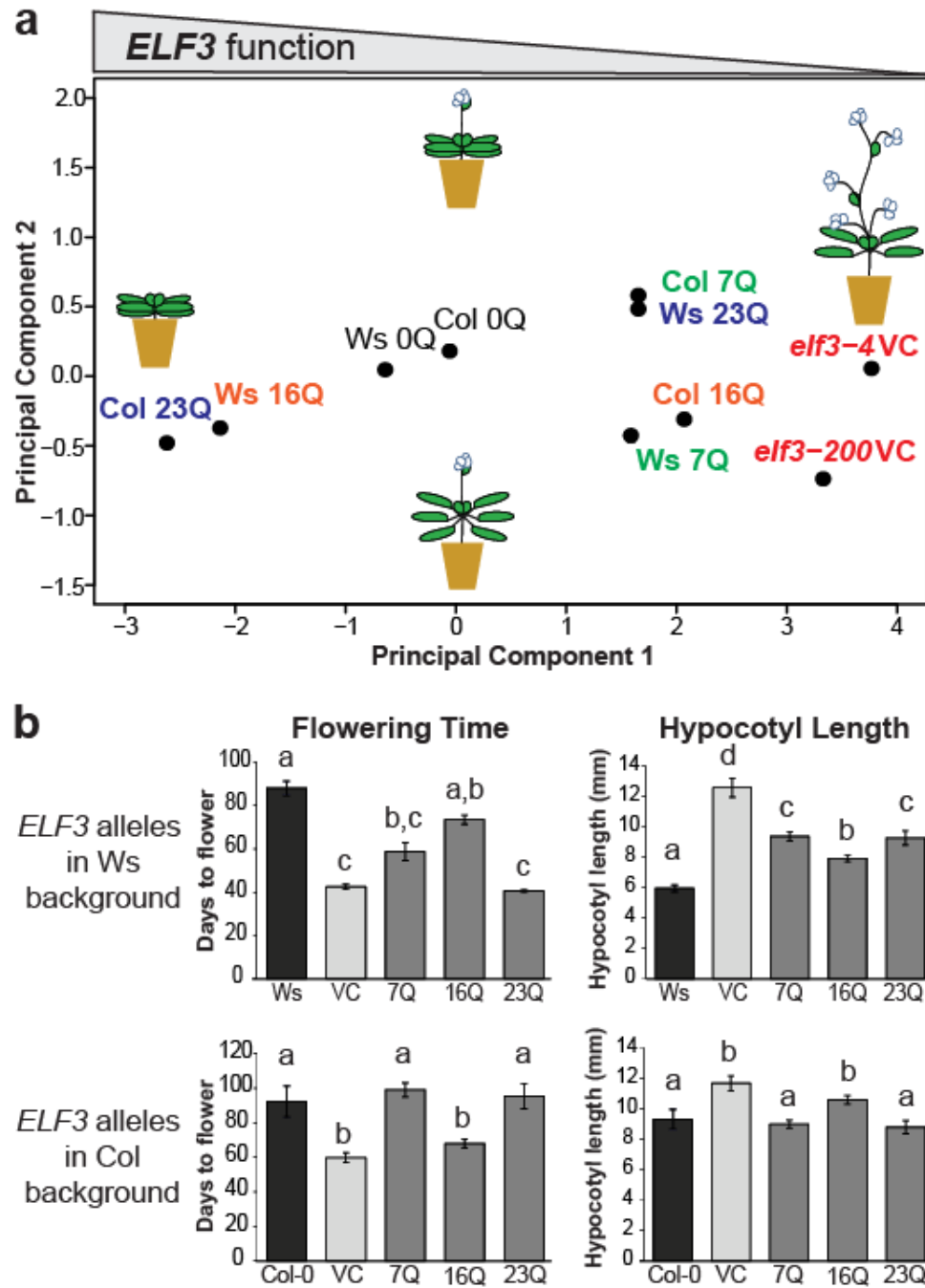
Figure 2.3

Figure 2.3: The phenotypic effects of ELF3-TR variation are strongly background-dependent. (A) PCA of developmental traits of all ELF3-TR alleles in Ws and Col genetic backgrounds. Shared color indicates a given ELF3-TR allele in both genetic backgrounds. A. thaliana images are as in Fig. 1A. The contributions of phenotypes to principal components are similar to Fig. 1A, except that PC2 is inverted (no effect on interpretation, loadings in Fig. A5C). Representative TR copy number alleles are shown from an analysis including all alleles (for all alleles see Fig. A5; for Col-background specific PCA, see Fig. A6). (B) Days to flower under SD and hypocotyl length under LD differ for particular TR alleles between Ws (Upper) and Col (Lower) backgrounds. ELF3-TR alleles are indicated with the number of Qs encoded, Ws and Col-0 are wild-type, elf3-4 and elf3-200 are respective vector controls (VC). Error bars represent SEM. Genotypes labeled with different letters differed significantly in phenotype by Tukey-HSD test. For all Col-background phenotype data, see Fig. S6 A☐G. Data are from multiple independently generated expression-matched (Fig. A1C and D) T3 and T4 lines for each TR copy number allele (Tables A2, A3). These experiments were repeated at least once with similar results.

2.7 Col **ELF3** *allele is not haploinsufficient in Col x Ws hybrids.*

To address whether Ws and Col-specific background effects are sufficient for altered hybrid phenotypes, we generated F1 populations between wild-type and elf3 null plants in the Ws and Col backgrounds and measured ELF3 function by assessing hypocotyl length. Ws x Col F1 hybrids resembled their wild-type parents (Figure 4, top). F1 hybrids containing both loss-of-function alleles had significantly longer hypocotyls than either parent. Both ELF3 alleles were haplosufficient in F1 crosses within their native backgrounds, as expected for recessive mutants . In stark contrast, we observe that ELF3-Col, but not ELF3-Ws, phenocopied the extreme hypocotyl length of the double loss-of-function mutant. Consistent with the results from our transgenic lines, our

F1 hybrid data suggest that full ELF3 function depends on a permissive genetic background.

Unfortunately, propagation of these F_1 hybrids to the F_2 generation and subsequent Col x Ws crosses revealed that these data do not generalize to other crosses (Figure 4 bottom, Figure A7), and probably represent a spontaneous mutation in the Col background leading to ELF3 inactivation. Repetition of the experiment with newly generated F_1s led to inconsistent results, and propagation of various batches to the F_2 generation (Figure A7) supported a mutation linked to the *ELF3* locus in the Col parent underlying apparent haploinsufficiency. In the face of such equivocal evidence, we suggest that more intensive genetic or biochemical experiments will be necessary to determine the relevant background modifiers of *ELF3-TR* variation. For such approaches, refer to Chapter 5.

2.8 Discussion

Our results demonstrate that natural ELF3 polyQ variation that is not associated with disease has dramatic phenotypic consequences, and that these consequences depend on genetic background. For ELF3, in at least the Ws background, the relationship between TR copy number and phenotype does not follow a linear or threshold pattern as observed for other coding TR and polyQ disorders (1, 2, 16, 17, 20). Studies correlating TR variation with phenotype often apply linear models, treating TR copy number as a quantitative variable (19, 22, 23). Our data show that this approach is not appropriate for all TRs. Instead, *ELF3-TR* alleles seem "matched" to specific genetic backgrounds, in which they are functional, whereas they are incompatible with other backgrounds. Variable TRs, and the *ELF3-TR* in particular, have been previously suggested as agents of adaptation to new external environments (1, 16, 17, 20, 24, 44). Our results, demonstrating strong background effects, suggest that polyQ-encoding TRs are also agents of coadaptation within genomes. We speculate that the observed background effects arise from background-specific polymorphisms in genes encoding physically interact-
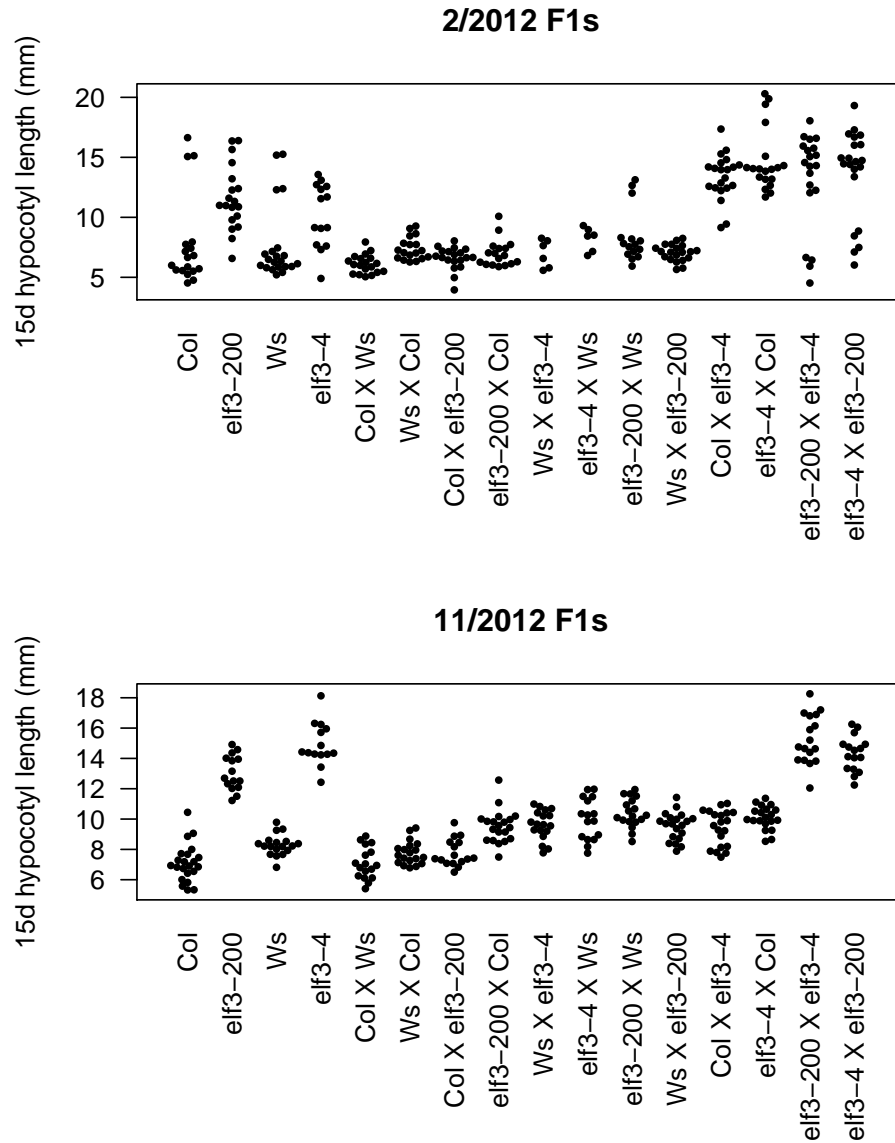
Figure 2.4: Top: ELF3-Col appears haploinsufficient in a hybrid Col x Ws genetic background. Bottom: Haploinsufficiency disappears in new seed batches. Hypocotyl length under SD was measured in seedlings from parental and F1 lines. elf3-ws is the elf3-4 loss-of-function mutant in the Ws background; elf3-col is the elf3-200 loss-of-function mutant in the Col background. Reciprocal crosses for each F1 showed similar results. n ≥ 15 for each genotype, except for Ws x elf3-ws in the top panel (n = 5). Each experiment was repeated with similar results (drawing from the same seed batch).

ing proteins (26, 28-30). TRs have a far higher mutation rate than non-repeated regions (10^{-4} per site per generation for TR vs. 10^{-8} for SNPs) (45, 46) and, as we show, their expansion or contraction can have dramatic phenotypic impact. ELF3's partner proteins may have acquired compensatory mutations to accommodate new *ELF3-TR* variants and vice versa. Alternative explanations for the background effects are compensatory mutations in ELF3 (intragenic suppressors), or ELF3 interactions that are unique to a given background. Intragenic variation and protein modification can play an important role in polyQ-mediated phenotypes (47, 48). At least for the ELF3-Col allele, however, our F1 data are not consistent with intragenic suppressors. Consistent with polyQ-mediated background effects, in at least one case, a modifier mutation has been shown to delay onset of Huntington's disease (11). Hypothetically, population genetic approaches could identify incompatible alleles that may contribute to variable disease onset in patients with polyQ expansions and to ELF3-dependent background effects in A. thaliana. However, the great diversity of TR alleles compared to SNP alleles and the small number of individuals carrying specific TR alleles render a population genetics approach infeasible. Extensive genetic mapping or other experimental approaches will be needed to identify the determinants of *ELF3-TR* dependent background effects. As TRs are rapidly evolving, we speculate that polyQ-mediated incompatibilities and the resulting fitness loss in hybrids and their offspring may contribute to disruption of gene flow between closely related populations. This speciation mechanism would be of particular importance for organisms with many polyQ-encoding TRs, thousands of offspring, and an inbreeding life style. Even in humans, however, about 1% of proteins contain polyQ tracts (13, 14, 45). Our results identify TR copy number variation, and in particular polyQ variation, as a phenotypically important class of genetic variation that warrants genome-wide assessment in model organisms, crops, and humans alike.

Chapter 3

THE CONSERVED *PFT1* TANDEM REPEAT IS CRUCIAL FOR PROPER FLOWERING IN *ARABIDOPSIS THALIANA*

A version of this chapter was published under the following reference:

Pauline Rival, Maximilian O. Press, Jacob Bale, Tanya Grancharova, Soledad F. Undurraga, and Christine Queitsch. The Conserved PFT1 Tandem Repeat is Crucial for Proper Flowering in *Arabidopsis thaliana. Genetics*, 198(2):747-754, August 2014.

Pauline Rival, Jacob Bale, Tanya Grancharova, and Soledad Undurraga contributed transgenic lines and experiments.

Supporting figures and tables can be found in Appendix B.

3.1 Summary

It is widely appreciated that short tandem repeat (STR) variation underlies substantial phenotypic variation in organisms. Some propose that the high mutation rates of STRs in functional genomic regions facilitate evolutionary adaptation. Despite their high mutation rate, some STRs show little to no variation in populations. One such STR occurs in the Arabidopsis thaliana gene PFT1 (MED25), where it encodes an interrupted polyglutamine tract. Though the PFT1 STR is large (270 bp), and thus expected to be extremely variable, it shows only minuscule variation across A. thaliana strains. We hypothesized that the PFT1 STR is under selective constraint, due to previously undescribed roles in PFT1 function. We investigated this hypothesis using plants expressing transgenic PFT1 constructs with either an endogenous STR or with synthetic STRs of varying length. Transgenic plants carrying the endogenous PFT1 STR gener-

ally performed best in complementing a pft1 null mutant across adult PFT1-dependent traits. In stark contrast, transgenic plants carrying a PFT1 transgene lacking the STR phenocopied a pft1 loss-of-function mutant for flowering time phenotypes, and were generally hypomorphic for other traits, establishing the functional importance of this domain. Transgenic plants carrying various synthetic constructs occupied the phenotypic space between wild-type and pft1-loss-of-function mutants. By varying PFT1 STR length, we discovered that PFT1 can act as either an activator or repressor of flowering in a photoperiod-dependent manner. We conclude that the PFT1 STR is constrained to its approximate wild-type length by its various functional requirements. Our study implies that there is strong selection on STRs not only to generate allelic diversity, but also to maintain certain lengths pursuant to optimal molecular function.

3.2 *Introduction*

Short tandem repeats (STRs, microsatellites) are ubiquitous and unstable genomic elements that have extremely high mutation rates (Subramanian et al. 2003; Legendre et al. 2007; Eckert and Hile 2009), leading to STR unit number variation within populations. STR variation in coding and regulatory regions can have significant phenotypic consequences (Gemayel et al. 2010). For example, several devastating human diseases, including Huntington☐s disease and spinocerebellar ataxias, are caused by expanded STR alleles (Hannan 2010). However, STR variation can also confer beneficial phenotypic variation and may facilitate adaptation to new environments (Fondon et al. 2008; Gemayel et al. 2010). For example, in Saccharomyces cerevisiae natural polyQ variation in the FLO1 protein underlies variation in flocculation, which is important for stress resistance and biofilm formation in yeasts (Verstrepen et al. 2005). Natural STR variants of the Arabidopsis thaliana gene ELF3, which encode variable polyQ tracts, can phenocopy elf3 loss-of-function phenotypes in a common reference background (Undurraga et al. 2012). Moreover, the phenotypic effects of ELF3 STR variants differed dramatically between the divergent backgrounds Col and Ws, consistent with the exis-

tence of background-specific modifiers. Genetic incompatibilities involving variation in several other STRs have been described in plants, flies, and fish (Peixoto et al. 1998; Scarpino et al. 2013; Rosas et al. 2014). Taken together, these observations argue that STR variation underlies substantial phenotypic variation, and may also underlie some genetic incompatibilities. The A. thaliana gene PHYTOCHROME AND FLOWER-ING TIME 1 (PFT1, MEDIATOR 25, MED25) contains an STR of unknown function. In contrast to the comparatively short and pure ELF3 STR, the PFT1 STR encodes a long (90 amino acids in PFT1, vs. 7-29 for ELF3), periodically interrupted polyQ tract. The far greater length of the PFT1 STR leads to the prediction that its allelic variation should be greater than that of the highly variable ELF3 STR (Legendre et al. 2007, http://www.igs.cnrs-mrs.fr/TandemRepeat/Plant/index.php). However, in a set of diverse A. thaliana strains, PFT1 STR variation was negligible compared to that of the ELF3 STR (Supp. Table 1). Also, unlike ELF3, the PFT1 polyQ is conserved in plants as distant as rice, though its purity decreases with increasing evolutionary distance from A. thaliana. A glutamine-rich C-terminus is conserved even in metazoan MED25 (File S1). Recent studies of coding STRs suggested that there may be different classes of STR. Specifically, conserved tandem repeats appear in genes with substantially different functions from genes containing non-conserved tandem repeats (Schaper et al. 2014). Consequently, PFT1/MED25 polyQ conservation may functionally differentiate the PFT1 STR from the ELF3 STR. PFT1 encodes a subunit of Mediator, a conserved multi-subunit complex that acts as a molecular bridge between enhancer-bound transcriptional regulators and RNA polymerase II to initiate transcription (Bäckström et al. 2007; Conaway and Conaway 2011). PFT1/MED25 is shared across multicellular organisms but absent in yeast. In A. thaliana, the PFT1 protein binds to at least 19 different transcription factors (Elfving et al. 2011; Ou et al. 2011; Çevik et al. 2012; Chen et al. 2012) and has known roles in regulating a diverse set of processes such as organ size determination (Xu and Li 2011), ROS signaling in roots (Sundaravelpandian et al. 2013), biotic and abiotic stress (Elfving et al. 2011; Kidd et al. 2009; Chen et

al. 2012), phyB-mediated-light signaling, shade avoidance and flowering (Cerdⁿn and Chory 2003; Wollenberg et al. 2008; Iⁿigo, Alvarez, et al. 2012; Klose et al. 2012). PFT1 was initially identified as a nuclear protein that negatively regulates the phyB pathway to promote flowering in response to specific light conditions (Cerdⁿn and Chory 2003; Wollenberg et al. 2008). Recently, Iⁿigo and colleagues (2012) showed that PFT1 activates CONSTANS (CO) transcription and FLOWERING LOCUS T (FT) transcription in a CO-independent manner. Specifically, proteasome-dependent degradation of PFT1 is required to activate FT transcription and to promote flowering (Iⁿigo, Giraldez, et al. 2012). The wide range of PFT1-dependent phenotypes is unsurprising given its function in transcription initiation, yet it remains poorly understood how PFT1 integrates these many signaling pathways. Given the conservation of the PFT1 polyQ tract and the known propensity of polyQ tracts for protein-protein and protein-DNA interactions (Escher et al. 2000; Schaefer et al. 2012), we hypothesized that this polyQ tract plays a role in the integration of multiple signaling pathways and is hence functionally constrained in length. We tested this hypothesis by generating transgenic lines expressing PFT1 with STRs of variable length and evaluating these lines for several PFT1-dependent developmental phenotypes. We show that the PFT1 STR is crucial for PFT1 function, and that PFT1-dependent phenotypes vary significantly with the length of the PFT1 STR. Specifically, the endogenous STR allele performed best for complementing the flowering and shade avoidance defects of the pft1-2 null mutant, though not for early seedling phenotypes. Our data indicate that most assayed PFT1-dependent phenotypes require a permissive PFT1 STR length. Taken together, our results suggest that the natural PFT1 STR length is constrained by the requirement of integrating multiple signaling pathways to determine diverse adult phenotypes.

3.3   Methods

3.3.1   Cloning

A 1000 bp region directly upstream of the PFT1 coding region was amplified and cloned into the pBGW gateway vector (Karimi et al. 2002) to create the entry vector pBGW-PFT1p. A full-length PFT1 cDNA clone, BX816858, was obtained from the French Plant Genomic Resources Center (INRA, CNRGV), and used as the starting material for all our constructs. The PFT1 gene was cloned into the pENTR4 gateway vector (Invitrogen) and the repeat region was modified by site-directed mutagenesis with QuikChange (Agilent Technologies), followed by restriction digestions and ligations. The modified PFT1 alleles were finally transferred to the pBGW-PFT1p vector via recombination using LR clonase (Invitrogen) to yield the final expression vectors. Seven constructs expressing various polyQ lengths (Table B2), plus an empty vector control, were used to transform homozygous pft1-2 mutants by the floral dip method (Clough and Bent 1998). Putative transgenics were selected for herbicide resistance with Basta (Liberty herbicide; Bayer Crop Science) and the presence of the transgene was confirmed by PCR analysis. Homozygous T3 and T4 plants with relative PFT1 expression levels between 0.5 and 4 times the expression of Col-0 were utilized for all experiments described. A minimum of two independent lines per construct was used for all experiments.

3.3.2   Expression analysis

All protocols were performed according to manufacture□s recommendations unless otherwise noted. Total RNA was extracted from 30mg of 10-days-old seedlings with the Promega SV Total RNA Isolation System (Promega). 2 ?g of total RNA were subjected to an exhaustive DNaseI treatment using the Ambion Turbo DNA-free Kit (Life Technologies). cDNA was synthesized from 100-300 ng of DNase-treated RNA samples with the Roche Transcriptor First Strand cDNA Synthesis Kit (Roche). Quanti-

tative Real-Time PCR was performed in a LightCyler®480 system (Roche) using the 480 DNA SYBR Green I Master kit. Three technical replicates were done for each sample. RT-PCR was performed under the following conditions: 5 min at 95 °C, followed by 35 cycles of 15 s at 95 °C, 20 s at 55 °C, and 20s at 72 °C. After amplification, a melting-curve analysis was performed. Expression of UBC21 (At5g25760) was measured as a reference in each sample, and used to calculate relative PFT1 expression. All expression values were normalized relative to WT expression, which was always set to 1.0. To measure splice forms, the protocol was the same but reactions were carried out in a standard thermal cycler and visualized on 2% agarose stained with ethidium bromide. For primers, see Table B4.

3.3.3 *Plant Materials and Growth Conditions*

Homozygous plants for the T-DNA insertional mutant SALK 129555, pft1-2, were isolated by PCR analysis from an F2 population obtained from the Arabidopsis Stock Center (ABRC) (Alonso et al. 2003). Plants were genotyped with the T-DNA specific primer LBb1 (`http://signal.salk.edu/tdna_FAQs.html`) and gene-specific primers (Table B4). Seeds were stratified at 4oC for 3 days prior to shifting to the designated growth conditions, with the shift day considered day 0. For flowering time experiments, plants were seeded using a randomized design with 15-20 replicates per line in 4x9 pot trays. Trays were rotated 1800 and one position clockwise everyday in order to further reduce any possible position effect. Plants for LD were grown in 16 hours of light and 8 hours of darkness per 24 hour period. Bolting was called once the stem reached 1 cm in height. Full strength MS media containing MES, vitamins, 1% sucrose, and 0.24% phytagar was used for hypocotyl experiments. For germination experiments, half-strength MS media was used, supplemented with 1% sucrose, 0.5 g/L MES, and 2.4 g/L phytagel containing 200 mM NaCl or H2O mock treatment with the pH adjusted to 5.7. All media was sterilized by autoclaving with 30 minutes of sterilization time. Seeds for tissue culture were surface sterilized with ethanol treatment prior to plat-

ing and left at 40C for 3 days prior to shifting to the designated growth conditions. Plants for hypocotyl experiments were grown with 16 hours at 22 °C and 8 hours at 200C in continuous darkness following an initial 2 hour exposure to light in order to induce germination. Germination experiments were scored on day 4 under LD at 20-22 °C. ImageJ software was utilized to make all hypocotyl and root length measurements. Raw phenotypic data are included as File S3.

3.3.4 Statistical analysis

All statistical analyses and plots were performed in R version 2.15.1 with ? = 0.05 (R Development Core Team 2012). Phenotypic data were analyzed using the analysis of variance (ANOVA), followed by Tukey□s HSD tests for the differences of groups within the ANOVA. Tukey□s HSD is a standard post-hoc test for multiple comparisons of the means of groups with homogeneous variance that corrects for the number of comparisons performed. Principal component analysis was performed using the prcomp() function after scaling each phenotypic variable to mean=0 and variance=1 across lines (phenotypes are not measured on the same quantitative scale; for example, SD flowering time ranges from 80 to 140 days, whereas LD rosette leaves ranges 5-15 leaves).

3.3.5 Sequence Analysis

Length of ELF3 and PFT1 STRs were determined by Sanger (dideoxy) sequencing. Raw sequencing data are available on the *Genetics* website (http://www.genetics.org/content/198/2/747.long). PFT1 and MED25 reference amino acid sequences were obtained from KEGG (Ogata et al. 1999) and aligned with Clustal Omega v1.0.3 with default options (Sievers et al. 2011).

3.4 Natural variation of PFT1 STR

We used Sanger sequencing to evaluate our expectation of high PFT1 STR variation across A. thaliana strains. However, we observed only three alleles of very similar size

(encoding 88, 89 and 90 amino acids, Table B1), in contrast to six different alleles of the much shorter ELF3 STR among these strains, some of which are three times the length of the reference allele (Undurraga et al. 2012). These data implied that the PFT1 and ELF3 STRs respond to different selective pressures. In coding STRs, high variation has been associated with positive selection (Laidlaw et al. 2007), though some basal level of neutral variation is expected due to the high mutation rate of STRs. We hypothesized that the PFT1 STR was constrained to this particular length by PFT1□s functional requirements. To test this hypothesis, we generated transgenic A. thaliana carrying PFT1 transgenes with various STR lengths in an isogenic pft1-2 mutant background. These transgenics included an empty vector control (VC), 0R, 0.34R, 0.5R, .75R, 1R (endogenous PFT1 STR allele), 1.27R, and 1.5R constructs. All STRs are given as their approximate proportion of WT STR length □ for instance, the 1R transgenic line contains the WT STR allele in the pft1-2 background (Table B2). We used expression analysis to select transgenic lines with similar PFT1 expression levels (Table B3).

3.5 *The PFT1 STR length is essential for wild-type flowering and shade avoidance*

We first evaluated the functionality of the different transgenic lines in flowering phenotypes. Removing the STR entirely substantially delayed flowering under long days (LD, phenotypes days to flower, rosette leaf number at flowering; Figure 1A). In LD, any STR allele other than 0R was able to rescue the pft1-2 late-flowering phenotype. Indeed, one allele (1.5R) showed earlier flowering than WT (Figure 1B, 1C), whereas other alleles provided a complete or nearly complete rescue of the pft1-2 mutant (Figure 1D). In short days (SD), we observed an unexpected reversal in rosette leaf phenotypes (compare SD and LD rosette leaves, Figures 1B, 1D). Rather than flowering late (adding more leaves) as in LD, the loss-of-function pft1-2 mutant appeared to flower early (fewer leaves at onset of flowering). Only the endogenous STR (1R) fully rescued this unexpected phenotype (Figure 1D). We observed the same mean trend for days to flowering

in SD, although differences were not statistically significant, even for pft1-2 (Figure 1D). This discrepancy may be due to insufficient power, or to a physiological decoupling of number of rosette leaves at flowering and days to flowering phenotypes in pft1-2 under SD conditions. Regardless, our results indicate that pft1-2□s late-flowering phenotype is specific to LD conditions. Our observation of this reversal in flowering time-related phenotypes appears to contradict previous data (Cerd□n and Chory 2003). However, a closer examination of this data reveals that the previously reported rosette leaf numbers in SD for the pft1-2 mutant show a similar trend. PFT1 STR length shows an approximately linear positive relationship with the SD rosette leaf phenotype, forming an allelic series of phenotypic severity. This allelic series strongly supports our observation of either slower growth rate (i.e. delayed addition of leaves) or early flowering of pft1-2 as measured by SD rosette leaves at flowering. PFT1 genetically interacts with the red/far-red light receptor phyB, which governs petiole length through the shade avoidance response (Cerd□n and Chory 2003; Wollenberg et al. 2008). We measured petiole length at bolting for plants grown under LD to evaluate the strength of their shade avoidance response, and thus whether the genetic interaction is affected by repeat length. Like the flowering time phenotypes, we found that the 1R allele most effectively rescued the long-petiole phenotype of the pft1-2 null among all STR alleles (Figure 2), though some alleles (e.g. 1.5R) show a rescue that is nearly as good. In summary, plants expressing the 1R transgene most closely resembled wild-type plants across a range of adult phenotypes. In contrast, the other STR alleles showed inconsistent performance across these phenotypes, rescuing only some phenotypes or at times out-performing wild-type.

3.6 *PFT1 STR alleles fail to rescue early seedling phenotypes*

We next assessed quantitative phenotypes in early seedling development, some of which had been previously connected to PFT1 function. Specifically, we measured hypocotyl and root length of dark-grown seedlings and examined germination in the presence of

salt (known to be defective in pft1 mutants) (Elfving et al. 2011). The pft1-2 mutant showed the previously reported effect on hypocotyl length as well as a novel defect in root length (Figure 3A). None of the transgenic lines, including the one containing the 1R allele, effectively rescued these pft1-2 phenotypes (Figure 3A). Similarly, 1R was not able to rescue the germination defect of pft1-2 on high-salt media. However, both the 1.5R and 0.5R alleles were able to rescue this phenotype (Figure 3B). In summary, no single STR allele, including the endogenous 1R, was consistently able to rescue the early seedling phenotypes of the pft1-2 mutant. One explanation for the failure of the endogenous STR (PFT1-1R) to rescue early seedling phenotypes is that the PFT1 transgene represents only the larger of two splice forms. The smaller PFT1 splice form, which we did not test, may play a more important role in early seedling development. To explore this hypothesis, we measured mRNA levels of the two splice forms in pooled 7-day seedlings grown under the tested conditions and various adult tissues at flowering in Col-0 plants. However, we found that both splice forms were expressed in all samples, and in all samples the larger splice form was the predominant form (data not shown). The possibility remains that downstream regulation or tissue-specific expression may lead to a requirement for the smaller splice form in early seedlings.

3.7 *Summarizing PFT1 STR function across all tested phenotypes*

Given the complex phenotypic responses to PFT1 STR substitutions, results were equivocal as to which STR allele demonstrated the most □wild-type-like□ phenotype across traits, as measured by its sufficiency in rescuing pft1-2 null phenotypes. To summarize the various phenotypes, we calculated the mean of each quantitative phenotype for each allele, and used principal component analysis (PCA) to visualize the joint distribution of phenotypes observed. All STR alleles were distributed between the pft1-2 null and wild-type (WT) in PC1, which was strongly associated with adult traits and represented a majority of phenotypic variation among lines (Figure 4). PC1 showed that 1R was the most generally efficacious allele for adult phenotypes. However, 1R showed

incomplete rescue in early seedling phenotypes such as hypocotyl length, which drove PC2. All STR alleles showed substantial rescue in adult phenotypes, and even the 0R allele showed a partial rescue in some phenotypes; however, rescue of early seedling phenotypes was generally poor for all alleles. The first principal component also captured our observation that the pft1-2 flowering defect reversed sign in SD vs. LD: according to Figure 4, SD and LD quantitative phenotypes are both strongly represented on principal component 1, but they show opposite directionality. We take this observation as support of this hitherto-unknown complexity in PFT1 function.

3.8 Discussion

STR-containing proteins pose an intriguing puzzle □they are prone to in-frame mutations, which in many instances lead to dramatic phenotypic changes (Gemayel et al. 2010). Although STR-dependent variation has been linked to adaptation in a few cases, the presence of mutationally labile STRs in functionally important core components of cell biology seems counterintuitive. PFT1, also known as MED25, is a core component of the transcriptional machinery across eukaryotes and contains an STR that is predicted to be highly variable in length. Contrary to this prediction, we found PFT1 STR variation to be minimal, consistent with substantial functional constraint. The existing residual variation (2% of reference STR length, as opposed to >100% for the ELF3 STR in the same A. thaliana strains) suggests that the PFT1 STR is mutationally labile like other STRs. In fact, several of the synthetic PFT1 alleles examined in this study arose spontaneously during cloning. Strong functional constraint, however, may select against such deviations in STR length in planta. Here, we establish the essentiality of the full-length PFT1 STR and its encoded polyQ tract for proper PFT1 function in A. thaliana. We found that diverse developmental phenotypes were altered by the substitution of alternative STR lengths for the endogenous length. Leveraging the support of the PFT1 STR allelic series, we report new aspects of PFT1 function in flowering time and root development.

3.8.1   *The PFT1 STR is required for PFT1 function in adult traits*

The PFT1 0R lines did not effectively complement pft1-2 for adult phenotypes, suggesting a crucial role of the PFT1 STR in regulating the onset of flowering and shade avoidance. Generally, PFT1-1R was most effective in producing wild-type-like adult phenotypes. The precise length of the STR, however, seemed less important for the onset of flowering in LD. With exception of PFT1-0R, all other STR alleles were also able to rescue the loss-of-function mutant to some extent, suggesting that as long as some repeat sequence is present, the PFT1 gene product can fulfill this function. Under other conditions, and for other adult phenotypes, requirements for PFT1 STR length appeared more stringent. Specifically, under SD, the rosette leaf number phenotype of the pft1-2 mutant can only be rescued by PFT1-1R, while STR alleles perform worse with increasing distance from this length □optimum□.

3.8.2   *pft1-2 mutants are late-flowering in LD but not SD*

pft1-2 plants had fewer rosette leaves at flowering in SD, but more rosette leaves in LD, consistent with previous, largely undiscussed observations (Cerd□n and Chory 2003). Under LD conditions, pft1 null mutants flowered late, as described in several previous studies (Cerd□n and Chory 2003; Wollenberg et al. 2008), but we observe no such phenotype under SD conditions, contradicting at least one prior study (Cerd□n and Chory 2003). These data suggest that while PFT1 functions as a flowering activator under LD, its role is more complex under SD. One recent study showed that PFT1 function in LD is dependent upon its ability to bind E3 ubiquitin ligases (I□igo, Giraldez, et al. 2012). Inhibition of proteasome activity also prevents PFT1 from promoting FT transcription and thus inducing flowering, suggesting that degradation of PFT1 or associated proteins is a critical feature of PFT1□s transcriptional activation of flowering in LD. If this degradation is somehow down-regulated in SD, PFT1 could switch from a flowering activator to a repressor, through decreased Mediator complex turnover at

promoters. Recent studies raised the possibility that different PFT1-dependent signaling cascades have different requirements for PFT1 turnover (Ou et al. 2011; Kidd et al. 2009), which may contribute to the condition-specific PFT1 flowering phenotype we observe. Conservatively, we conclude that the regulatory process that mediates the phenotypic reversal between LD and SD depends on the endogenous PFT1 STR allele, suggesting that the polyQ is crucial to PFT1□s activity as both activator and potentially as a repressor of flowering.

3.8.3 *Incomplete complementation of germination and hypocotyl length by the PFT1 constructs*

Whereas pft1-2 adult phenotypes were rescued by the PFT1-1R allele, most of our transgenic lines could not fully rescue pft1-2 early seedling phenotypes of 1) germination under salt, 2) hypocotyl length, and 3) root length. The PFT1 gene is predicted to have two different splice forms, the larger of which was used to generate our constructs (both splice forms contain the STR). Several studies have shown that, under stress conditions, different splice forms of the same gene can play distinct roles (Yan et al. 2012; Leviatan et al. 2013; Staiger and Brown 2013). We note that the conditions under which PFT1-1R fails to complement are also potentially stressful conditions (artificial media, sucrose, high salt, dark). The shorter splice form of PFT1 may be required in signaling pathways triggered under stress conditions. We presume that the failure to complement results from a deficiency related to this missing splice form. However, hypocotyl length was the only trait in which all examined STR alleles resembled the pft1-2 mutant. The significant functional differentiation among the STR alleles for root length and germination suggests that the large splice form does retain at least some function in early seedling traits.

3.8.4 *Implications for STR and PFT1 biology*

Implications for STR and PFT1 biology: Coding and regulatory STRs have been previously studied and discussed as a means of facilitating evolutionary innovation (Verstrepen et al. 2005). However, this means of innovation is based upon the same sequence characteristics that promote protein-protein and protein-DNA binding (Escher et al. 2000; Schaefer et al. 2012), such that STR variability must be balanced against functional constraints. This balance has recently been described for a set of 18 coding dinucleotide STRs in humans, which are maintained by natural selection even though any mutation is likely to cause frame-shift mutations (Haasl and Payseur 2014). These results, coupled with our observations, lend credence to these authors□ previous argument that not all STRs act as agents of adaptive change (Haasl and Payseur 2013). Considering again the possibility that more conserved coding tandem repeats have distinct functions from non-conserved tandem repeats (Schaper et al. 2014), we suggest that PFT1 and ELF3 can serve as models for these two selective regimes, and that the structural roles of their respective polyQs underlie the differences in natural variation between the two. In some cases, such as ELF3, high variability is not always inconsistent with function, even while holding genetic background constant (Undurraga et al. 2012). In PFT1, we have identified a STR whose low variability reflects strong functional constraints. We speculate that these constraints are associated with a structural role for the PFT1 polyQ in the Mediator complex, either in protein-protein interactions with other subunits or in protein-DNA interactions with target promoters. Given that a glutamine-rich C-terminus appears to be a conserved feature of MED25 even in metazoans (File S1), we expect that our results are generalizable to Mediator function wherever this protein is present. Future work will be necessary in understanding possible mechanisms by which the MED25 polyQ might facilitate Mediator complex function and contribute to ontogeny throughout life. Moreover, attempts must be made to understand the biological and structural characteristics unique to polyQ-

containing proteins that tolerate (or encourage) polyQ variation, as opposed to those polyQ-containing proteins (like PFT1) that are under strong functional constraints.

Chapter 4

SHORT TANDEM REPEATS AND QUANTITATIVE GENETICS

Portions of this chapter were published under the following references:

- Maximilian O. Press, Keisha D. Carlson, and Christine Queitsch.

  The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics*, 30(11):504-512, August 2014.

- Keisha D. Carlson, Peter H. Sudmant, Maximilian O. Press, Evan E. Eichler, Jay Shendure, and Christine Queitsch. MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Research*, 25(5):750-761, May 2015.

Keisha Carlson and Peter Sudmant contributed STR genotyping data. Supporting figures and tables can be found in Appendix C.

4.1 Abstract

Short tandem repeat (STR) variation has been proposed as a major explanatory factor in the heritability of complex traits in humans and model organisms. However, we still struggle to incorporate STR variation into genotype-phenotype maps. Here, we review the promise of STRs in contributing to complex trait heritability, and highlight the challenges that STRs pose due to their repetitive nature. We argue that STR variants are more likely than single nucleotide variants to have epistatic interactions, reiterate the need for targeted assays to accurately genotype STRs, and call for more appropriate statistical methods in detecting STR-phenotype associations. Lastly, somatic STR

variation within individuals may serve as a read-out of disease susceptibility, and is thus potentially a valuable covariate for future association studies.

4.2 The □missing heritability□ of complex diseases and STR-variation.

Complex diseases such as diabetes, various cancers, cardiovascular disease, and neurological disorders cluster in families, and are thus considered to have a genetic component [1□3] (Glossary). The identification of these genetic factors has proven challenging; although genome-wide association (GWA) studies have identified many genetic variants that are associated with complex diseases, these generally confer less disease risk than expected from empirical estimates of heritability. This discrepancy, termed the □missing heritability□, has been attributed to many factors [1□6]. A trivial explanation is that shared environments among relatives may artificially inflate estimates of heritability. However, missing heritability may also be due to variants in the human genome that are currently inaccessible at a population scale [1,2]. One such class of variation is short tandem repeat (STR) unit number variation. Some have previously suggested that adding STR variation to existing genetic models would considerably increase the proportion of heritability explained by genetic factors in human disease [7,8]. Three percent of the human genome consists of STRs [9] and *6% of human coding* regions are estimated to contain STR variation [10,11]. Recently, the first catalog of genome-wide population-scale human STR variation has appeared [12], opening up new possibilities for understanding the contribution of STRs to human genetic diseases. This catalog, and similar data sources [13], have appeared decades after initial calls for the assessment of the role of STRs in phenotypic variation [14], lagging behind surveys of other genomic elements. Much of the initial interest in STRs was generated by the discovery of phenomena such as genetic anticipation, which are mediated by the unique features of STRs [15]. As we will discuss, new and forthcoming data sources will help to realize the long-deferred promise of STRs for explaining heritability. STRs consist of short (2-10 bp) DNA sequences (units) that are repeated head-to-tail multiple times.

This structure causes frequent errors in recombination and replication that add or subtract units, leading to STR mutation rates that are 10-fold to 104-fold higher than those of non-repetitive loci [16,17]. Due to technical barriers, STR variation has until very recently remained inaccessible to genome-wide assessment. STRs are often conserved (even if their unit number or even sequence changes), especially in coding sequences [18□21]. In both humans and the yeast Saccharomyces cerevisiae, promoter regions are known to be dramatically enriched for STRs [22,23]. In coding regions, STRs tend to occur in genes with roles in transcriptional regulation, DNA binding, protein-protein binding, and developmental processes [16,21, 22]. These consistent functional enrichments across vastly diverged lineages suggest important functional roles for STRs. Indeed, analysis of STR variation in the Drosophila Genetic Reference Panel identified dozens of associations between STR variants and quantitative phenotypes in recombinant inbred fly lines [13]. Moreover, accumulating evidence from exhaustive genetic studies shows that STR variation has dramatic, often background-dependent phenotypic effects in model organisms [25□29]. Together, these findings suggest that STR variation has the potential to dramatically revise the heritability estimates attributable to genetic factors. The high STR mutation rate also leads to substantial somatic variation of STR loci within individuals. In fact, this somatic variation, also called microsatellite instability (MSI), has been used for decades as a biomarker for different classes of cancer [30]. Recent studies demonstrate that organisms exposed to various environmental stresses and perturbations show increased genome instability, including MSI [31□34]. MSI may be useful as a biomarker for cellular stress states that may predispose to disease. The broad interest in STR variation has led to the development of techniques for high-throughput genotyping of STRs [35,36] and an explosion of analysis tools for extracting STR variation from existing sequence data [37□39]. However, the precision of these methods remains limited, due to a combination of low effective coverage of STRs and the lack of robust models for distinguishing technical error from somatic variation. Attempts to use STR variation for GWA in a fashion equivalent to

SNV variation may be underpowered and confounded by the unique characteristics of this class of variants. In this review, we discuss the latest advances in these fields, and lay out a set of priorities for the future study of STRs.

4.3 STR variation is associated with human genetic diseases

Within coding regions, STR mutations are generally in-frame additions and subtractions of repeat units, resulting in proteins with variable, low-complexity amino acid runs [21]. These mutations can result in phenotypic effects and lead to genetic disorders; several neurological diseases (spinocerebellar ataxias, Huntington□s disease, spinobulbar muscular atrophy, dentatorubral-pallidoluysian atrophy, intellectual disability, etc.) are a consequence of dramatically expanded STR alleles [7,40,41]. Many of these disease-associated STR expansions behave as dominant gain-of-function mutations [7]. However, even comparatively modest coding STR variation may confer disease risk or behavioral phenotypes, according to a variety of single-marker association studies [42□45]; for instance, variants in separate coding STRs in RUNX2 are associated with defects in bone mineralization, higher incidence of fractures [46,47]; STR variation in this gene in dogs is also associated with craniofacial phenotypes [48]. Noncoding STR variation in regulatory sequences can affect transcription, RNA stability, and chromatin organization. For instance, certain STR variants alter CFTR expression and thus cystic fibrosis status [16]. We take these studies as evidence that STR variation, even in the absence of large expansions, may contribute significantly to the heritability of human traits and genetic diseases. The severity of the STR expansion-associated diseases may suggest that natural selection should eliminate STRs in functional regions, but several recent studies across many organisms indicate that variable STRs are globally maintained [19,20,24,49,50]. For example, the pre-expansion polyQ-encoding STR in the human gene SCA2 is under positive selection, suggesting that this variable STR is actively maintained in spite of the pathogenic expansions that do occasionally occur and cause spinocerebellar ataxia [51]. Considering both the evidence of

positive selection on STRs and the functional enrichments of STR-containing genes, several authors have proposed that functional STRs are maintained because they confer ☐evolvability☐, or the capacity for fast adaptation [21,22,52☐54]. This suggestion is intriguing, in part because many STR mutations are dominant, and, when beneficial, can quickly sweep to fixation. Although we do not further discuss these evolutionary considerations here, they underscore the phenotypic potential of STR variation.

4.4 STR variation has dramatic background-dependent effects on phenotype

To date, the functional consequences of unit number variation in selected STRs have been studied in plants, fungi, flies, voles, dogs, and fish [25,27,28,55☐57], among other organisms. In Saccharomyces cerevisiae, STR unit number in the FLO1 gene accurately predicts the phenotype of cell-cell and cell-substrate adhesion (flocculation); flocculation provides protection against various stresses [57,58]. STR variation in yeast promoters has been shown to alter gene expression [22]. In Drosophila melanogaster, Neurospora crassa, and Arabidopsis thaliana, natural coding STR variation in circadian clock genes alters diurnal rhythmicity and developmental timing [25☐27,59]. Some have proposed that the large phenotypic responses to selection observed in the Canidae are a consequence of elevated STR mutation rates relative to other mammalian clades [48,53]. We can state unambiguously that naturally variable STRs underlie dramatic phenotypic variation in model organisms. Beyond the observable fact that variable STRs affect phenotype, we can make specific predictions about the components of phenotypic variation that they affect. Both theoretical expectations and empirical data indicate that STR variants are likely to participate in epistatic interactions, and probably more so than most SNVs. One plausible hypothesis is that STRs act as mutational modifiers of other loci, as may be expected intuitively from their elevated mutation rate (Box 1, Figure I). This expectation is borne out in the handful of studies reporting exhaustive genetic analysis of STRs. For instance, in the Xiphophorus genus of

fish, a genetic incompatibility has recently been attributed to the interaction between the xmrk oncogene and an STR in the promoter of the tumor suppressor cdkn2a/b [29,60]. If the xmrk gene product is not properly regulated by cdkn2a/b, fish develop fatal melanomas, a two-locus Bateson-Dobzhansky-Muller incompatibility described in classic genetic experiments (Figure 1A) [61□63]. Expansions in the cdkn2a/b promoter STR are associated with the presence of a functional copy of the xmrk oncogene across species, and are thought to functionally repress the activity of the xmrk gene product through increased dosage of the tumor suppressor [29]. Similarly, we have shown that natural variation in the polyQ-encoding ELF3 STR significantly affects all ELF3-dependent phenotypes in the plant A. thaliana, with ELF3 STR length and phenotype showing a strikingly nonlinear relationship (Figure 1B)[25]. Some naturally occurring ELF3 STR variants phenocopy elf3-loss-function mutants in a common reference background (Figure 1B), suggesting background-specific modifiers. Indeed, when we compare the phenotypic effects of each ELF3 STR variant between two divergent backgrounds, Columbia (Col-0) and Wassilewskija (Ws), we find dramatic differences. The endogenous STR alleles from these two strains (Col-0 7 units, Ws 16 units) show mutual incompatibility when exchanged between backgrounds. The ELF3 protein is thought to function as an □adaptor protein□ or physical bridge in diverse protein complexes [64,65]. We speculated that background-specific polymorphisms in these interacting proteins underlie the ELF3 STR-dependent background effect. Also in A. thaliana, a variable STR in the promoter of the CONSTANS gene has been linked to phenotypic variation in the onset of flowering [28]. CONSTANS encodes a major regulatory protein that promotes flowering. Transgenic experiments demonstrate that this regulatory STR variation affects CONSTANS expression and hence onset of flowering. However, the effects of this STR variation depend on the presence of a functional allele of FRIGIDA, a negative regulator of flowering that is highly polymorphic across A. thaliana populations. A dramatic example of incompatibility can be found in an intronic repeat in the IIL1 gene in A. thaliana, which was found to be dramatically

expanded in one strain [55]. The expansion delayed flowering under high temperatures, but when crossed into the reference genetic background, a strongly interacting locus modifies this phenotype. In the Drosophila genus, coding STR variation in the per gene co-evolves with other variants [59,66]. Transgenic flies expressing chimeric per genes with a D. melanogaster STR domain fused to a D. pseudoobscura flanking region (and vice versa) have arrhythmic circadian clocks, indicating the modifying effect of flanking variation in generating an STR-based genetic incompatibility. Among STRs subjected to exhaustive genetic study, to our knowledge, only the yeast FLO1 coding STR has no known modifiers due to variation in genetic background [57]. In addition to these exhaustive genetic studies, there are several other observations that support the role of the genetic background in controlling the phenotypic effects of STRs. For instance, experiments in Caenorhabditis elegans and human cells indicate that the phenotypic effects of proteins with expanded polyQ tracts are modulated by genetic background [67], or by variants in interacting proteins [68]. In humans, genetic association studies indicate the existence of genetic modifiers of polyQ expansion disorders for both Huntington□s disease [69] and spinocerebellar ataxias [70]. Taken together, these experimental and observational data support our argument that functional STRs are likely to be enriched for variants in epistasis with other loci. STRs with background-dependent phenotypic effects tend to either encode polyQ tracts or reside in promoter regions. There are good reasons to expect that these STR classes might be enriched in DNA/protein-protein interactions that could underlie epistasis. PolyQ tracts, specifically, often bind DNA surfaces [71], and an analysis of human protein interactome data found that polyQ-containing proteins engage in more physical interactions with other proteins than those without polyQs [72]. Similarly, noncoding STRs in regulatory regions may compensate for mutations in trans-acting factors, as observed for the STRs in the cdkn2a/b promoter in Xiphophorus [29] and in the CONSTANS promoter in A. thaliana [28]. We suggest that polymorphisms in protein interaction partners or in transcriptional regulators are plausible explanations for the observed background ef-

fects. In summary, we expect that STR variation is likely to contribute a substantial epistatic component to heritability, which has important implications for their use in explaining phenotypic variation.

4.5 *Lack of statistical models for detecting STR-phenotype associations in GWA.*

Assuming that we obtain accurate, population-scale genotype data for STRs, we may not yet have statistical tools appropriate for detecting STR associations with phenotype [8]. In diploid organisms, a biallelic SNV is typically analyzed by modeling phenotype as a function of the number of non-reference alleles at that locus (0, 1, or 2) in each individual. A null hypothesis of no monotonic relationship between phenotype and the allele count is then formulated and tested [83]. This framework cannot accommodate more than two alleles, which we would expect for many STRs. Simply using tagged SNVs linked to STRs to perform GWA is unfeasible, because linkage disequilibrium decays very quickly between SNVs and STRs across human populations [12]. To address these complications, a previous study attempted GWA between STR genotypes and human disease phenotypes by comparing relative frequencies of various alleles in pooled DNA from cases and controls [84]. By pooling samples, this approach eases the analysis of multiallelic loci, but it loses information by ignoring specific individuals. In a more recent study, the authors used logistic regression and the analysis of variance to detect associations between STR alleles and quantitative phenotypes in an inbred Drosophila mapping population [13]. Given that significant associations were detected, such approaches may be sufficiently powerful in recombinant inbred lines. However, their strategy relied on homozygosity, and considered multiallelic STRs in a pairwise fashion, so these straightforward methods will lose power with outbred populations and multiallelic STRs. The central confounder of these studies is that most STRs of appreciable variability (and thus, interest) are multiallelic, as a simple consequence of the STR mutational mechanism [17]. This multiallelic feature could be accommodated
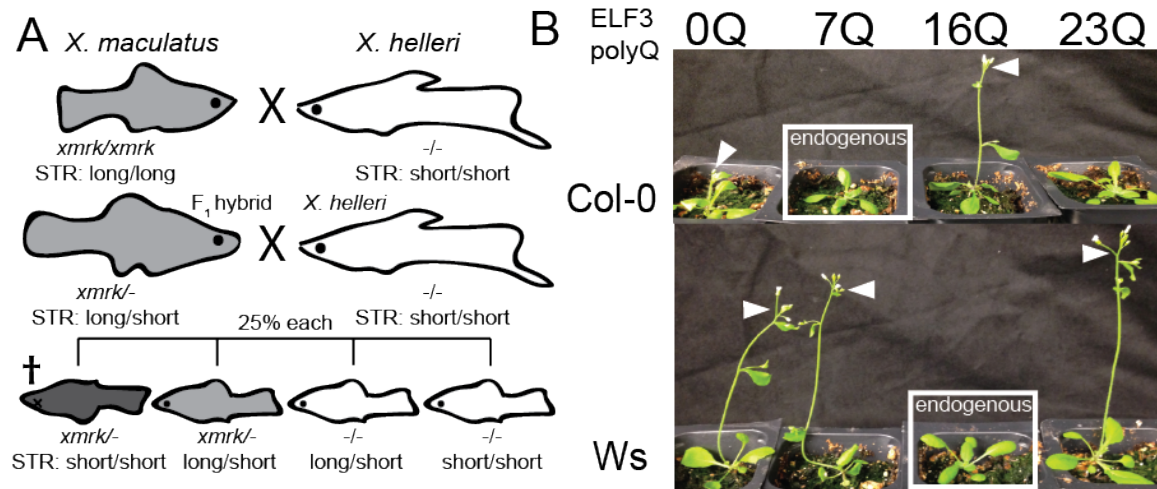
Figure 4.1: Genetic and transgenic analysis reveals STR-mediated incompatibilities. A, the Gordon-Kosswig-Anders cross shows a genetic incompatibility between two fish species in the Xiphophorus genus. Modified from Meierjohann and Schartl [63]. F1 hybrids back-crossed to their X. helleri parent yield a 3:1 ratio of viability, where the inviables result from co-segregation of the functional xmrk gene and a short STR allele in the cdkn2a/b promoter. Shading indicates melanism conferred by xmrk. B, genetic background is epistatic to effects of ELF3 STR variation in A. thaliana. Expression-matched transgenic plants with various alleles of the ELF3 STR in the Columbia (Col-0) and Wassilewskija (Ws) backgrounds, showing endogenous, exogenous, and synthetic (□o□) alleles in each background [25]. White boxes indicate transgenic plants carrying the ELF3 STR endogenous to their respective background; white arrowheads indicate early-flowering ELF3 STR genotypes (elf3 mutants and poorly-functioning ELF3 STR alleles confer early flowering).

by treating STR alleles categorically, but this choice entails a corresponding reduction in power, because many alleles are rare. Some studies have reported linear associations between STR unit number and quantitative phenotypes [27,57], suggesting that using simple tests of linear correlation between these variables may be a powerful option. However, this linearity (or even monotonicity) of the relationship between STR unit number genotype and phenotype is a poorly-supported assumption [25]. Nonetheless, STR unit number is a numerical variable, and it would be preferable to gain power from treating it as such. For instance, more similar STR unit number genotypes might be associated with more similar phenotypes, but this intuition may be difficult to generalize. Lastly, both intuition (Figure C.2)) and the studies discussed above lead us to expect that relatively many phenotypically relevant variable STRs will show epistasis with other loci. This epistasis will reduce power in tests of association between STRs and phenotype [85], given the inadequacy of the current paradigm of quantitative genetics in detecting and modeling the effects of epistasis [85,86]. At present, targeted and exhaustive genetic studies (as described above) are the only effective method for understanding the effects of epistasis. In total, these obstacles present a daunting challenge for the integration of STR genotypes into the current genotype-phenotype maps. Overall, we call for a reappraisal of statistical methodologies for use in GWA with STR variation to account for these various STR-specific confounders.

4.6 *Modifier mutations leading to epistasis are expected in STRs.*

We have previously proposed that STRs might be more susceptible to genetic interactions [25], as we will briefly explicate here. Consider a simple two-locus haploid model under panmixis, in which loci A and B each start with a single allele (ab) and have the same probability p per generation of mutating to a second allele (a* or b*), with p also as the probability per generation of reverting mutations (Figure I). Let us further assume that A and B are in sign epistasis [99] (that is, a*b and/or ab* have fitness less than ab and a*b*). To escape the unfavorable a*b genotype, the organism may either revert to

ab or mutate forward to a*b*. When the A and B loci have equal mutation rates, we expect that the reversion of a single mutant is just as likely as a second mutation, and consequently that a*b* individuals will appear only relatively rarely and slowly. However, consider a similar model, in which locus B has an elevated mutation rate pb > pa. In this case, the a*b genotype has a higher probability of a second, modifying mutation to a*b* than of a reversion to ab. Moreover, flux along the other mutational path (ab ? ab* ? a*b*) will be increased. In sum, a*b* genotypes will arise at higher rates, and will attain their equilibrium frequency much more rapidly, if either A or B has an elevated mutation rate [100] (p.131). This scenario can lead quickly to an equilibrium population in which incompatible epistatic alleles are frequent, even though recombinants have lower fitness. Relaxing the assumption of no population structure will further speed this process. Consequently, we would expect STRs and other loci with high mutation rates to be more likely to modify other alleles than loci with lower mutation rates, as long as we assume that all loci are equally capable of genetic interactions. This process may be referred to as □coadaptation□. For a rigorous model of the evolution of hybrid incompatibility, see Orr [101].

Figure I. A locus with higher mutation rates allows genetic modification of unfavorable genotypes at interacting loci. Top, a model of evolution under epistasis with only one slow mutation rate. Middle, a model of evolution under epistasis with a slow and a fast mutation rate. Boxes represent loci, stars represent SNV-type mutations, black and white checkering indicates an STR locus (a/b, a*/b, and a*/b* signify different genotypes). Arrows with numbers represent possible mutations and their respective rates. Bottom, fitness of each genotype under both models. We expect that the model with two mutation rates will occupy the fully derived state (a*/b*) more quickly.

4.7 Association of STR-variation with phenotypic characters in A. thaliana.

4.8 Analysis of selection on STR-variation in A. thaliana.

4.9 Concluding remarks

The study of STRs and other under-ascertained genomic elements has the potential to reshape our model of the heritability of complex diseases and traits, both in terms of the overall proportion of heritability explained, and in terms of the components of heritability themselves (Outstanding Questions). Experimental studies in model organisms have taught us that the phenotypic effects of genome-wide STR variation are both dramatic and impossible to understand without taking epistasis into account. In the future, our understanding will be improved by 1) accurate STR population-scale genotyping, 2) more appropriate statistical methods for analyzing STR-phenotype associations, and 3) a broader description of epistasis between STR variation and other loci in determining phenotype.

Chapter 5

THE VARIABLE ELF3 POLYGLUTAMINE HUBS AN EPISTATIC NETWORK

Supporting figures and tables can be found in Appendix C.

Chapter 6

ELF3 POLYGLUTAMINE VARIATION REVEALS A PIF4-INDEPENDENT ROLE IN THERMORESPONSIVE FLOWERING

Supporting figures and tables can be found in Appendix D.

Chapter 7

GENOME-SCALE CO-EVOLUTIONARY INFERENCE IDENTIFIES FUNCTIONS AND CLIENTS OF BACTERIAL HSP90

A version of this chapter was published under the following reference:

Maximilian O. Press, Hui Li, Nicole Creanza, Guenter Kramer, Christine Queitsch, Victor Sourjik, and Elhanan Borenstein. Genome-scale co-evolutionary inference identifies functions and clients of bacterial Hsp90. *PLoS Genetics*, 9(7):e1003631, 2013.

Hui Li, Guenter Kramer, and Victor Sourjik contributed E. coli strains, experiments, and figures. Nicole Creanza and Aviv Regev contributed ideas and preliminary analyses at the initiation of this work.

Supporting figures and tables can be found in Appendix E.

7.1 Abstract

The molecular chaperone Hsp90 is essential in eukaryotes, in which it facilitates the folding of developmental regulators and signal transduction proteins known as Hsp90 clients. In contrast, Hsp90 is not essential in bacteria, and a broad characterization of its molecular and organismal function is lacking. To enable such characterization, we used a genome-scale phylogenetic analysis to identify genes that co-evolve with bacterial Hsp90. We find that genes whose gain and loss were coordinated with Hsp90 throughout bacterial evolution tended to function in flagellar assembly, chemotaxis, and bacterial secretion, suggesting that Hsp90 may aid assembly of protein complexes. To add to the limited set of known bacterial Hsp90 clients, we further developed a statistical method to predict putative clients. We validated our predictions by demon-

strating that the flagellar protein FliN and the chemotaxis kinase CheA behaved as Hsp90 clients in E. coli, confirming the predicted role of Hsp90 in chemotaxis and flagellar assembly. Furthermore, normal Hsp90 function is important for wild-type motility and/or chemotaxis in E. coli. This novel function of bacterial Hsp90 agreed with our subsequent finding that Hsp90 is associated with a preference for multiple habitats and may therefore face a complex selection regime. Taken together, our results reveal previously unknown functions of bacterial Hsp90 and open avenues for future experimental exploration by implicating Hsp90 in the assembly of membrane protein complexes and adaptation to novel environments.

7.2   *Introduction*

In eukaryotes, the universally conserved and essential chaperone Hsp90 aids the folding of key proteins in development and responses to environmental stimuli [1□3]. In yeast, up to 10% of all proteins are estimated to be Hsp90 clients under standard culture conditions [4]. Hsp90 function is even more important under stressful conditions that challenge protein folding, such as increased temperature [5]. The activity of eukaryotic Hsp90 is further modulated by various co-chaperones, which confer substrate specificity and alter protein folding kinetics [2,5]. Depletion of eukaryotic Hsp90 in vivo increases phenotypic variation, reveals □cryptic□ heritable variation, and increases penetrance of mutations [6□9]. Accordingly, eukaryotic Hsp90 enables organisms to maintain a stable phenotype in the face of environmental and genetic perturbation and to correctly interpret environmental stimuli. In stark contrast, in prokarya, Hsp90 is not essential [10] and many bacterial genomes lack Hsp90 altogether [11]. Among Archaea, only very few species contain Hsp90, and those are thought to have gained Hsp90 horizontally from bacteria [11,12]. This fragmented phylogenetic pattern likely results from multiple independent gains and losses, though phylogenetic reconstructions are confused by ancient Hsp90 paralogy [11,12]. At the amino acid level, the E. coli Hsp90 (High-temperature protein G or HtpG) is 42% identical to its human ho-

molog, suggesting strong stabilizing selection consistent with functional conservation [13]. Indeed, E. coli Hsp90 appears to retain generic protein chaperone activity [14] and homologous mutations cause chaperone defects in both the prokaryotic E. coli and eukaryotic yeast Hsp90 [15]. However, there are no identified obligate Hsp90 co-chaperones in bacteria, adding to the uncertainty regarding the extent of its client spectrum and specificity. To date, only three proteins have been implicated as Hsp90 clients in bacteria, which have non-overlapping functions in ribosome assembly, the assembly of light-harvesting complexes, and the CRISPR/Cas immunity system [16□18]. Several other proteins have been shown to physically interact with the chaperone [19,20]. These data, together with our knowledge on eukaryotic Hsp90 function, have given rise to the speculation that Hsp90 may facilitate the assembly of oligomeric protein complexes in bacteria, much like it does in eukaryotes [21]. Unlike in eukaryotes, however, further exploration of Hsp90□s functional role in bacteria has proven challenging because there are no pleiotropic Hsp90-dependent phenotypes. To address this challenge, we used a genome-scale co-evolutionary □guilt-by-association□ approach [22,23] to explore the spectrum of conserved Hsp90-associated genes, functions, and organismal traits. Hsp90-associated genes tended to function in flagellar assembly, chemotaxis, and secretion. Consistent with these functions, Hsp90-associated organismal traits included the ability to inhabit multiple environments. To add to the sparse list of known bacterial Hsp90 clients, we further developed a statistical method to predict putative Hsp90 clients, which included flagellar, ribosomal, and chaperone proteins. We validated our predictions experimentally, focusing on two candidates functioning in motility and chemotaxis. Indeed, both the flagellar protein FliN and the kinase CheA were found to be Hsp90 clients in vivo. Our findings demonstrate the power of co-evolutionary inference to correctly identify substrates and functions of conserved genes like bacterial Hsp90.

7.2.1 Results

7.2.2 Hsp90 paralogs in bacteria

Our method for inferring the function of bacterial Hsp90 is based on analysis of its distribution across the bacterial phylogeny. However, this analysis is complicated by the existence of multiple ancient Hsp90 paralogs in bacteria. These paralogs may be older than existing phyla in bacteria [11,12], and may have evolved distinct functions on this enormous time scale. To address this issue and to identify each paralog, we first clustered bacterial Hsp90s by sequence identity. We identified 897 bacterial Hsp90 protein sequences in the KEGG database [24] and built a neighbor-joining gene tree of bacterial Hsp90s (Figure E1A-B). We observed two well-supported long-branching clades as well as several less confident divisions in the tree (Figure E1B). These two long-branching clades contain sequences corresponding to the □hsp90B□ and □hsp90C□ paralogs that were described previously [11,12]. All other branches correspond to □hsp90A□ [11], which is the largest of the Hsp90 families in bacteria (Figure E1C, Supporting Text). Notably, hsp90A is the lineage out of which all eukaryotic Hsp90s (excluding mitochondrial and chloroplast Hsp90s) are derived. Moreover, the E. coli gene htpG belongs to the hsp90A family, and its gene product is the best-studied bacterial Hsp90 protein. For these reasons, we restricted our analysis to hsp90A.

7.2.3 Genome-wide detection of genes co-evolving with hsp90A

We set out to identify orthologous groups whose presence and absence profiles across bacterial species are associated with the presence and absence profile of hsp90A. To avoid spurious associations, any such comparative analysis must go beyond a na□ve comparison of presence/absence patterns across genomes and incorporate phylogenetic information [25]. To this end, we used BayesTraits [26□28], a computational framework for phylogenetic analysis of character evolution. Given the states (e.g., presence/absence) of two characters across some set of species and a phylogenetic tree re-

lating these species, BayesTraits evaluates the likelihood of various evolutionary models throughout the tree. This approach can be utilized, for example, to determine whether these two characters evolve in a mutually dependent vs. an independent fashion. We used BayesTraits to detect associations between hsp90A and 4646 other orthologous groups in bacteria (which hereafter we shall refer to as □genes□ for simplicity). We used the tree constructed by Ciccarelli et al. [29] as a model phylogeny (Figure 1). In this initial analysis, we tested for any kind of dependency between hsp90A and other genes, and did not make specific assumptions about the nature of the relationship between hsp90A and the genes in question [28]. Specifically, we compared a model in which the rate of gain and loss of a given gene is independent of the rate of gain and loss of hsp90A (independent evolution) vs. a model in which the rate of gain and loss of this gene is affected by the presence or absence of hsp90A or vice-versa (co-evolution). In total, we found 327 genes that co-evolve with hsp90A. We will refer to this set as hsp90A-associated genes. These hsp90A-associated genes were significantly enriched for annotations related to the flagellum and to bacterial secretion systems (Table 1). Moreover, out of the 16 hsp90A-associated bacterial secretion genes, 10 were part of the non-flagellar Type III secretion system, suggesting that hsp90A is associated specifically with this system rather than with secretion systems in general. Using a different and markedly more extensive phylogeny [30] provided similar results (see Supporting Text, Table S1), as did a pruned Ciccarelli tree without the species containing the hsp90B or hsp90C (see Supporting Text).

7.2.4 *Characterization of co-evolutionary dynamics*

The associations of hsp90A with other genes identified above are agnostic to the specific nature of the dependency between hsp90A and the gene in question. For example, our initial analysis could not distinguish between a positive association (i.e. genes tend to be gained and lost together) and a negative association (i.e. genes tend not to co-occur in genomes). Similarly, this analysis did not distinguish between genes whose gains and

losses are affected by the presence of hsp90A (but that do not themselves affect hsp90A evolution) and genes that exhibit mutually dependent dynamics with hsp90A. Without a quantitative estimate of the effects that hsp90A and its co-evolving partners have upon one another, inference of Hsp90A function and its relationship with other genes is challenging. To characterize the specific nature of the dependency between hsp90A and hsp90A-associated genes, we therefore examined rates of gain and loss inferred by BayesTraits. We focused on the two major non-overlapping hsp90A-associated functional categories, flagellar assembly and bacterial secretion. Considering, for example, fliI, a representative flagellar gene, we found that its gain and loss was strongly affected by the presence of hsp90A. Specifically, in the presence of hsp90A, fliI was often gained and rarely lost, whereas it was rarely gained and often lost when hsp90A is absent (Figure 2A). This pattern was common to all hsp90A-associated flagellar genes (Figures 2C, S2), suggesting a positive association between hsp90A and flagellar genes throughout evolution. In contrast, the co-evolutionary relationship between hsp90A and yscN, a representative nonflagellar type III secretion system gene, was markedly different, with yscN presence strongly affecting the gain and loss of hsp90A (Figure 2B). Specifically, the presence of yscN was associated with a large increase in the rates of gain and (even more dramatically) loss of hsp90A relative to these rates in its absence. Again, this pattern was common to all hsp90A-associated bacterial secretion genes (Figures 2D, S3, S4), suggesting a negative association between hsp90A and nonflagellar secretion genes throughout evolution. To further validate the fundamentally distinct co-evolutionary dynamics of these two groups of genes, we considered four different co-evolutionary models: (1) hsp90A and the gene in question are independent (null); (2) hsp90A and the gene in question are mutually dependent; (3) hsp90A is dependent on the gene in question but not vice versa, and (4) the gene in question is dependent upon hsp90A but not vice versa (Methods). We used the Akaike Information Criterion (AIC [31]) to determine which of these 4 models best fit the co-evolutionary dynamics of each hsp90A-associated gene. As expected, none of the hsp90A-associated genes fit the independent

model. Of the 27 hsp90A-associated flagellar genes, 25 were classified as being dependent on hsp90A but not vice-versa (model 4). Of the 16 hsp90A-associated secretion system genes, 10 genes were classified as mutually dependent with hsp90A (model 2; 6 of which were Type III secretion system genes), whereas 6 were classified as affecting the evolution of hsp90A (model 3). Furthermore, considering all hsp90A-associated genes, we found that genes that best fit each of the evolutionary dependency models above (models 2, 3, and 4) were enriched for different functions (Table 1). Specifically, among genes dependent on hsp90A, flagellar motility was strongly enriched, whereas among genes mutually dependent on hsp90A, secretion system components were enriched. Taken together, these patterns suggest that flagellar genes and secretion system genes had markedly different regimes of co-evolution with hsp90A.

7.2.5 *Prediction of Hsp90A clients*

Although many genes exhibited distinct patterns of co-evolution with hsp90A, these patterns could be the result of indirect evolutionary relationships rather than the outcome of a direct interaction with Hsp90A. We therefore aimed to predict specific genes that encode putative hsp90A clients. Our method is based on the assumption that strong, conserved clients should be heavily dependent on Hsp90A, and thus should be found only rarely in the absence of hsp90A throughout evolution. To estimate the expected frequency of each hsp90A-associated gene with and without hsp90A, we used the inferred BayesTraits rates to calculate the steady-state probabilities of each of the 4 possible presence/absence states (Methods). These probabilities represent the proportion of the time that some arbitrary bacterial lineage will spend in each of the presence/absence states throughout evolution. From these probabilities we calculated a Putative Client Index (PCI) for each hsp90A-associated gene to evaluate how often it was present without hsp90A throughout evolution, compared to a null expectation (see Methods). This index is close to zero for genes that were infrequently present without hsp90A and were hence likely to be Hsp90A clients. We defined the genes with the

lowest PCI values as putative clients (Table 2; see also Supporting Text).

7.2.6 *Novel and known functions of putative Hsp90 clients*

Consistent with our prior analysis, several flagellar genes behaved as potential clients (Table 2). In particular, our set of putative clients included several genes (fliH, fliI, fliN) whose products had been previously shown to physically interact with Hsp90A in E. coli [19]. The products of these genes are cytoplasmic components of the flagellar rotor and export apparatuses. In contrast, nonflagellar type III secretion genes were all absent from the list of potential clients. In fact, nonflagellar type III secretion system components were rated as some of the least likely clients by our index (Figure 3). This disparity in predicted client status mirrors the different evolutionary relationships of these complexes with hsp90A (Figure 2). Chaperone/proteases (e.g. ClpA and PpiD) also ranked high in our list of potential clients. Hsp90A is known to collaborate with other chaperone systems such as DnaK [14,32] but to date no obligate co-chaperones have been described. The identified chaperone/proteases may represent such co-chaperones or collaborating chaperone systems, since our index cannot discriminate between Hsp90 clients and Hsp90 co-chaperones (or other collaborating proteins). Alternatively, these observed associations could simply indicate that components of the cytoplasmic stress response are dependent upon Hsp90A. We also found several unexpected putative clients, such as the 3-hydroxybutyryl-CoA dehydrogenase PaaH and the transcription termination factor Rho, which we predict to be the two strongest clients. Further study will be necessary to understand these associations and the underlying cause of the co-evolutionary association between these genes and hsp90A.

7.2.7 *Swimming motility and chemotaxis assays of Hsp90A-defective E. coli*

Our putative clients and the predicted chaperone role of Hsp90A in flagellar assembly are consistent with previous observations. Specifically, the deletion of E. coli hsp90A,

also known as htpG, resulted in reduced surface swarming movement [33]. We also previously observed physical interactions between the HtpG protein and certain flagellar proteins [19]. Yet, these observations lacked a clear demonstration of client status or mechanism, and E. coli swarming is a complex behavior that depends on numerous factors in addition to flagellar function [34]. We therefore set out to test our hypothesis that Hsp90A is physiologically important for flagellar assembly and function and that flagellar components are indeed Hsp90A clients. We examined the swimming motility phenotype of $\Delta htpG$ E. coli strains on soft-agar plates (Methods). In contrast to surface swarming, swimming is a less complex behavior, in which bacteria use functional flagella and chemotaxis components to swim from an inoculation point through agar pores, following nutrient gradients that are created by nutrient depletion within the colony. The soft-agar assay is routinely used to assay bacterial swimming motility and chemotaxis. To enhance our ability to detect differences between wild-type and $\Delta htpG$ cells, the assays were performed competitively. Competitive assays emphasize small differences between strains and reduce experimental error, thereby increasing the sensitivity of the assay. After mixing equal amounts of YFP-labeled WT and CFP-labeled $\Delta htpG$ strains, this mixture was inoculated in the center of a soft-agar plate and incubated at 34¡C for 8 hrs. We then counted cells of each strain in the plate center vs. the outer edge using fluorescence microscopy (Figure 4A). $\Delta htpG$ mutants migrated less efficiently to the plate□s outer edge relative to WT, confirming that they are partially deficient in their motility and/or chemotaxis (Figure 4B). This defect is apparently subtle, since little difference between WT and $\Delta htpG$ cells was observed in a non-competitive assay (Figure E5), but it could be revealed due to strong selection for cells with optimal motility and chemotaxis at the outer edge of the spreading bacterial population. We also tested the phenotype of the HtpG(E34A) mutant, which has reduced rates of ATP hydrolysis and is deficient in substrate refolding [14,35]. Since HtpG ATPase activity is necessary for release of clients, HtpG(E34A) is less efficient at releasing clients [36□38]. Indeed, this mutant showed stronger motility/chemotaxis defects than the $\Delta htpG$ strain (Fig-

ure E5), presumably due to sequestration of its client proteins. We therefore employed the HtpG(E34A) mutant in all subsequent assays as a more sensitive test of HtpG involvement. Taken together, our observations suggest that the motility defect may be due to the improper function or sequestration of HtpG clients.

7.2.8   *FRET observation of HtpG interactions with flagellar motor components*

To further investigate the in vivo interaction of HtpG with flagellar components, we used htpG-yfp and htpG(E34A)-yfp constructs expressed in WT cells to perform acceptor photobleaching FRET between HtpG and FliN-CFP over an E. coli growth curve. Motility of E. coli is known to increase at the transition from the early exponential to post-exponential phase of growth [39], and this experimental design enabled us to examine the HtpG-FliN interaction in the context of the flagellar assembly process. If HtpG is indeed involved in the assembly process of these structures, the interaction of HtpG with FliN should correspond temporally to the timing of flagellar assembly. Indeed, we found that the interaction with FliN peaked at OD600 = 0.2 (Figure 5A) and correlated well with the onset of cell motility in wild-type cells (Figure 5B). Moreover, the interaction of HtpG(E34A) with FliN was stronger and delayed compared to the binding of wild-type HtpG. This is consistent with the delayed release of clients by HtpG(E34A). Correspondingly, the onset of motility was delayed in cells expressing HtpG(E34A) (Figure 5B). These findings suggest that HtpG□s role in motility derives from a direct involvement in flagellar complex assembly. Given that both bacterial and eukaryotic Hsp90s are known to collaborate with Hsp70 in refolding proteins [14,40□42], we considered the possibility that this was also the case for bacterial flagellar assembly. We previously showed that some flagellar motor components interact with DnaK, the E. coli Hsp70 homolog [19]. Therefore, we repeated the FRET experiments testing for interactions between HtpG or HtpG(E34A) and FliN in a $\Delta cbpA\Delta dnaJ$ background. CbpA and DnaJ are DnaK co-chaperones and are essential for DnaK□dependent refolding activity [14]. DnaK should not be able

to pass substrates to HtpG in this mutant background. Indeed, we found that FRET interactions with FliN disappear for both HtpG proteins in this background (Figure E6A), suggesting that DnaK-dependent remodeling precedes HtpG action in flagellar complex assembly.

7.2.9   *FRET observation of HtpG interactions with chemoreceptor components*

Since a recent high-throughput assay showed kinases to be overrepresented among eukaryotic Hsp90 clients [43,44], we next examined whether the HtpG-dependent defects in chemotaxis may also be due to defective chemoreceptor kinase activity. Although no chemotaxis proteins were found in our list of the strongest putative clients, we did observe a significant enrichment of these components in the hsp90A-associated set (Table 1). We thus tested interactions between six chemoreceptor cluster components and HtpG(E34A) using, as before, acceptor photobleaching FRET (Table S4). We observed a strong interaction of HtpG(E34A) with the chemoreceptor kinase CheA. Our results suggest that the interactions between FliN and HtpG and CheA and HtpG are direct and do not depend on other flagellar or chemotaxis proteins, since these interactions are robust to deletion of flhC, which ablates expression of all endogenous flagellar and chemotaxis genes (Table S4) [19]. Moreover, the CheA dimerization domain was required for association with HtpG, supporting the hypothesis that HtpG aids oligomerization of its clients [17,45]. Testing HtpG interactions with other chemotaxis proteins of E. coli revealed an additional strong interaction with the dimeric phosphatase CheZ but not with other proteins (Table S4). We again examined the temporal dynamics of these interactions. Due to the hierarchical order of flagellar and chemotaxis gene expression [39,46], the assembly of chemoreceptor clusters is delayed compared to the assembly of flagellar motors as non-motile cells transition into motile cells. Indeed, the interaction of HtpG with CheA peaked at OD600 = 0.3, after the FliN peak (Figure 5A). Just as for FliN, the interaction of HtpG(E34A) with CheA was stronger and delayed compared to wild-type HtpG, and the HtpG-CheA interaction

disappeared in a $\Delta cbpA\Delta dnaJ$ background (Figure E6B). Collectively, these findings suggest that HtpG plays an important role in the assembly of both the flagellar motor and chemoreceptor clusters through separate client interactions.

7.2.10   *Association of hsp90A with life history traits in Bacteria*

Given the role of HtpG in chaperoning proteins that mediate interactions with the environment, and the known role of eukaryotic Hsp90 in phenotypic robustness, we finally examined whether hsp90A directly co-evolved with certain bacterial organismal traits. We considered several organismal traits, including aerobism, thermophilicity, halophilicity, the ability to form endospores, pathogenicity, motility, and habitat preferences (see Methods). We used BayesTraits and the Ciccarelli tree to identify traits that co-evolve with hsp90A. Out of the 11 analyzed traits, 4 exhibited significant associations with hsp90A ($p < 0.05$; Table S5), with the strongest association observed between hsp90A and the capacity to inhabit multiple habitats. Moreover, examining the gain and loss rates obtained, we found that hsp90A is gained and lost at significantly higher rates in organisms that inhabit multiple habitats (with no gains inferred in single habitat organisms), suggesting that a preference for multiple habitats imposes a different selection regime on hsp90A (Figure 6). We also tested whether the co-evolutionary dependency between hsp90A and multiple-habitat preferences was unidirectional, as we observed for some hsp90A-associated genes. Comparing the four co-evolutionary models described above and applying AIC to identify the best-fitting model, we found that hsp90A gain and loss depended on habitat preference, but not vice versa. This observation suggests that in organisms inhabiting multiple environments hsp90A is subjected to dynamically shifting selective pressures, potentially alternating between selection for and against hsp90A.

7.3 *Discussion*

We set out to discover Hsp90 functions conserved throughout the bacterial tree of life. We found that hsp90A, the most common paralog of bacterial Hsp90, bore strong signatures of co-evolution with several hundred genes and with specific life history traits, shedding light on its function and impact on evolutionary history. Most notably, we found that hsp90A co-evolved with membrane protein complexes such as flagella and other Type III secretion (T3S) systems. Our results suggest that Hsp90□s role in sensing and responding to environmental stimuli is conserved between bacteria and eukaryotes. Similar to verified eukaryotic Hsp90 clients [5], our predicted putative Hsp90A clients were a diverse group of proteins (e.g. the flagella protein FliN, the chaperone ClpA, and the ribosomal protein RluB; see Table 2) that tended to belong to specific functional categories (e.g. flagellar proteins, chaperones, and ribosomal components). As our methods can only infer associations between genes that are frequently gained and lost, we may substantially underestimate the number of hsp90A-associated genes and clients. However, the non-essentiality and frequent loss of hsp90A throughout bacterial diversity argues that genes not captured in our analysis (since they are not frequently gained and lost) are unlikely to be strongly dependent on the chaperone. The subtlety of the bacterial Hsp90 mutant phenotypes that we (and others) report implies that Hsp90□s role in cellular physiology has diverged between eukaryotes and prokaryotes [17,45,47]. In other words, either essential pieces of cellular physiology changed, or Hsp90 function changed. We favor the first hypothesis, because Hsp90 is well-conserved among bacteria, archaea, and humans at the sequence level [13], and retains a similar quaternary structure [48] and biochemical activity [15,37,44]. In contrast, bacterial and archaeal cells differ significantly from eukaryotic cells. Eukaryotic cells have higher cell compartmentalization, longer and multifunctional proteins with multiple domains [49], and increased protein interactome complexity [50]. Together with the existence of many eukaryotic Hsp90 co-chaperones, all these features may contribute

to the greater essentiality of Hsp90 in eukaryotes. The dependence of HtpG-client interactions upon the DnaK chaperone system, as observed by us and by others [14,15], argues that Hsp90A is well-integrated with other chaperone systems. Our putative clients included ClpA, the substrate adaptor for the ClpAP/ClpAXP chaperone/protease complexes, and PpiD, a periplasmic chaperone [51]. Like HtpG, PpiD is necessary for optimal swarming motility [33], suggesting that it may participate in flagellar assembly. We speculate that these proteins act as Hsp90A co-chaperones in some bacteria; alternatively, their dependence on Hsp90A may represent an example of collaborating chaperone systems. The best-characterized Hsp90 client in bacteria is the structural ribosomal protein L2 [15,18], which is near-universally conserved throughout life (and hence not detectable by our method). In addition to L2, other ribosomal proteins were found to interact with HtpG in large-scale proteomics analyses. In agreement with these observations, we found the ribosomal proteins RlmE and RluB among the predicted hsp90A clients. Although these chaperone and ribosomal proteins were predicted to be stronger clients than flagellar proteins, our experimental validation focused on the latter as their client status was suggested by previous observations [19,33]. We present four lines of evidence for HtpG client status for the flagellar protein FliN and the chemoreceptor kinase CheA, including direct interactions with HtpG, physiologically relevant timing of HtpG-FliN/CheA interactions, phenotypic consequences of reduced HtpG function in CheA/FliN-dependent traits, and dependence of CheA/FliN interactions with HtpG upon the Hsp40-Hsp70 pathway. The identification of FliN and CheA as HtpG clients is consistent with the hypothesis that bacterial Hsp90 facilitates the assembly of large membrane-associated protein complexes [17,45]. Curiously, whereas the flagellar T3S system contained Hsp90A clients, the nonflagellar T3S system is predicted to have an antagonistic relationship with Hsp90A. Nonflagellar T3S systems and the flagellar T3S systems are closely related (NF-T3SS and F-T3SS) [52,53]. 9 NF-T3SS components are directly homologous to flagellar components, of which 8 were found to co-evolve with hsp90A in our analysis. Yet, these 8 genes are predicted

to co-evolve antagonistically with hsp90A (Figure 3), whereas their flagellar homologs are mostly predicted to be clients (for instance, the fliI and yscN genes shown in Figure 2 are homologous). This result suggests that some relationship with Hsp90A is conserved between the two T3S systems, but with apparently opposite effects in each system. This result may reflect the fact that each of these systems is an adaptation to different ecological challenges. Specifically, we have shown that Hsp90A is important for flagella-enabled motility and chemotaxis in E. coli. This mode of motility is strongly adaptive in certain physical environments [34,54,55], and thus Hsp90A is likely to be associated with fitness in these environments through flagellar assembly. The presence of NF-T3SS is likewise an adaptation to certain biotic environments [55,56]. Our observation that organisms inhabiting multiple habitats experience fluctuating selection for hsp90A is also consistent with competing selection pressures. Representative genes of these homologous T3S families were not significantly associated with habitat preferences, arguing that hsp90A□s association with habitat preferences is not a byproduct of associations with T3S systems. We suggest that these two T3S systems constitute a link between Hsp90A and phenotypic robustness across different environments. Inferring function from evolutionary associations has some caveats. For instance, F-T3S systems can be found in genomes that lack hsp90A. If F-T3S systems include Hsp90A clients, then what may render Hsp90A-dependent stabilization dispensable in some bacteria? Experimental validation will be necessary to answer such questions, and to distinguish true client relationships from indirect co-evolutionary associations. As discussed before, our method is subject to gene set bias, in that only genes that are gained and/or lost frequently will have enough statistical power to reject the null hypothesis. Similarly, as our method assumes that relationships are maintained throughout the analyzed phylogeny, we cannot reliably detect genes that are associated with hsp90A in some organisms but not in others. Although much work remains to articulate the precise mechanistic relationships between hsp90A and its co-evolving genes, our results highlight the tremendous potential of evolutionary inference for guiding experimental

research. More generally, our study provides a successful example of how evolutionary perspectives and phylogenetic analyses can inform and advance the study of complex biological systems and the inference of elusive biological functions.

7.4 Methods

7.4.1 Prokaryotic Hsp90 paralogs

We downloaded all Hsp90 amino acid sequences (including all paralogs) for bacteria with full KEGG genome annotations from the KEGG database [24,57]. We aligned these sequences using ClustalO [58], and used the PHYLIP package [59] to construct neighbor-joining trees and assess their phylogenetic support through bootstrapping. We assigned Hsp90 families to branches according to bootstrap support for the branch and previous classifications [11,12].

7.4.2 Genome data

We acquired presence/absence patterns of genes across organisms from the KEGG database release 60.0 (in the form of KEGG Orthology/KO profiles) [57], and functional annotations from KEGG Class. Genes that were either present in fewer than five species or absent in fewer than five species in the tree of interest were dropped from our analysis, as these genes are unlikely to show meaningful signatures of co-evolution by this method.

7.4.3 Phylogenetic trees

We obtained the tree constructed by Ciccarelli et al. (Ciccarelli tree) [29] and pruned it to 148 bacterial species for which KEGG genome data was available. We also obtained the LTP104 version of the 16S/23S rRNA tree from the All-Species Living Tree Project (Yarza tree) [30,60]. We used ARB [61] to prune this tree to bacterial species for which KEGG genome data was available. We further pruned this tree to omit clades

placed paraphyletically at the taxonomic levels of phylum, class, order, and family. This filtered tree included 797 bacterial species. As BayesTraits cannot process trees with zero-length branches, all branch lengths equal to zero were replaced with a negligible branch length (0.00001, approximately an order of magnitude smaller than the next smallest branch length in each tree).

7.4.4 *Organismal trait data*

We acquired organismal trait data from the NCBI Entrez genome project, November 2011 [62]. We recoded all traits into presence/absence patterns for the trait in question. For instance, an organism found to be pathogenic towards any other organism was coded as '1' for the trait of pathogenicity, whereas an annotated organism that was never found to be pathogenic was coded as '0'. Similarly, we coded both thermophilic and hyperthermophilic organisms as '1' for the trait of thermophilicity, whereas all other annotated organisms were coded as '0'; anaerobic organisms were coded as □0□ for the trait of aerobicity, whereas all other annotated organisms were recoded as '1'. We define as inhabiting multiple habitats any organism that inhabits more than one of NCBI□s habitat categories. For BayesTraits analysis, the tree was pruned to include only species annotated for the trait in question (each trait analysis was accordingly performed on a slightly different set of species; see Table S5 for details on species number for each analysis).

7.4.5 *Detecting evolutionary associations with BayesTraits*

A complete description of the BayesTraits (v1.0) framework can be found elsewhere [26]. Briefly, consider a character with 2 states, 0 and 1. If a species has 2 such distinct characters, it can occupy 4 possible states: 1:(0,0), 2:(0,1), 3:(1,0), and 4:(1,1). Specifically, if these 2 characters represent the presence or absence of two genes, hsp90A and gene X, these four states correspond to (hsp90A-, X-), (hsp90A+, X-), (hsp90A-, X+), and (hsp90A+, X+). Evolution is then the process by which these genes are gained and lost

over time. Consider accordingly an evolutionary process where only one character can change state at a time. Such a process can then be described by 8 parameters for the rates of transition per unit time between these 4 states: Q = [q12, q13, q21, q31, q24, q34, q42, q43], where qxy is the rate of transition from state x to state y. BayesTraits implements this model of evolution as a continuous-time Markov process and estimates each of these rate parameters by maximum-likelihood (ML). We further validated that these ML-based rates are consistent with reversible-jump Markov chain Monte Carlo-derived estimates (Methods; Supporting Text). This estimation is based on a phylogeny and on the states of the two characters at the tips of the phylogeny. Having estimated these rates, BayesTraits additionally calculates the likelihood of the model based on the character states at the tips of the phylogeny. We can further compare different models of evolution by forcing certain parameters to be equal. We specifically considered the following 4 models:

1. hsp90A and X are independent (Q: q12 = q34, q21 = q43, q13 = q24, q31 = q42; 4 parameters total)

2. hsp90A and X are mutually dependent (No parameter restrictions; 8 parameters total)

3. X depends on hsp90A but not vice versa (Q: q12 = q34, q21 = q43; 6 parameters total)

4. hsp90A depends on X but not vice versa (Q: q13 = q24, q31 = q42; 6 parameters total)

7.4.6   *Identifying hsp90A-associated genes*

We used discrete from the BayesTraits package [26□28] to infer associations between hsp90A and other bacterial genes and between hsp90A and various organismal traits.

We first tested for an evolutionary association with hsp90A by comparing model 1 to model 2 above with a likelihood ratio test (LRT), as previously described [28]. In our likelihood-ratio tests, the 2Log(LR) approximates a X^2 test statistic for rejecting the independent model as a null hypothesis, and is calculated as twice the difference of the log-likelihoods of a co-evolutionary model and a model of evolutionary independence. The set of genes for which model 2 is preferred (i.e., model 1 is rejected as a null hypothesis) have an evolutionary association with hsp90A. Since different runs of the BayesTraits maximum likelihood method can potentially produce different parameter values, we repeated this procedure 100 times, each potentially resulting in a different gene set. We validated that these sets are similar and the choice of gene set does not substantially affect downstream analysis (Supporting Text). Any gene that was found to be associated with hsp90A in at least 90 runs was defined as hsp90A-associated gene. See Supporting Text for more details.

7.4.7   *Reversible-jump Markov chain Monte Carlo analysis*

We selected 10 genes at random from the hsp90A-associated set and used the BayesTraits implementation of reversible-jump Markov chain Monte Carlo to estimate the rate parameters for their gain and loss in concert with hsp90A [63]. For each of these 10 genes, we used an exponential rate prior with mean and variance equal to 30, and ran the chain for 150 million iterations while sampling every 100 iterations. We discarded the first 75 million iterations as burn-in and used the remaining iterations as a posterior distribution of rate parameter estimates. We used Tracer v1.5 [64] and previously described criteria to evaluate chain convergence in this remaining sample [65]. For each rate, we used the median of its posterior distribution in this sample as a point estimate.

7.4.8   *Co-evolutionary model selection*

To provide an accurate description of the co-evolutionary dynamics of hsp90A-associated genes, we further applied BayesTraits to these genes, estimating the likelihood of each

of the four models described above. We identified the best fit model for each gene using the Akaike Information Criterion (AIC) [31], taking into account both the likelihood score and the number of parameters in each model. We again repeated this procedure 100 times and classified a gene into a specific co-evolutionary model only if it fit this same model in at least 90 runs (see Supporting Text for more details). This two stage scheme, first identifying associated genes and then selecting a model that best describes their evolutionary relationship with hsp90A, provides a more stringent test of co-evolution and supports a simple approach for multiple testing correction.

7.4.9 Prediction of Hsp90A clients in bacteria

We used BayesTraits-derived evolutionary transition rates under the fully unrestricted model to estimate residence times in specific states (for instance, the proportion of time spent by bacteria in a state where both hsp90A and some other gene are present, vs. the time when only the other gene is present) under steady state dynamics. For a given gene, the probability of being in one of the four states, A: (hsp90A absent, Gene absent), B: (hsp90A present, Gene absent), C: (hsp90A absent, Gene present), D: (hsp90A present, Gene present) at a very small increment of timeΔt after time t is given by:

$$A_{t+\Delta t} = A_t - (q12 + q13) * A_t\Delta t + (q21 * B_t)\Delta t + (q31 * C_t)\Delta t + (0 * D_t)\Delta t$$

$$A_{t+\Delta t} = B_t + (q12 * A_t)\Delta t - (q21 + q24) * B_t\Delta t + (0 * C_t)\Delta t + (q42 * D_t)\Delta t$$

$$A_{t+\Delta t} = C_t + (q13 * A_t)\Delta t + (0 * B_t)\Delta t - (q31 + q34) * C_t\Delta t + (q43 * D_t)\Delta t$$

$$A_{t+\Delta t} = D_t + (0 * A_t)\Delta t + (q24 * B_t)\Delta t + (q34 * C_t)\Delta t - (q43 + q42) * D_t\Delta t$$

We can differentiate this to obtain the instantaneous change in each probability:

$$dA/dt = (q12 + q13) * A_0 + (q21 * B_0) + (q31 * C_0) + (0 * D_0)$$

$$dB/dt = (q12 * A_0) - (q21 + q24) * B_0 + (0 * C_0) + (q42 * D_0)$$

$$dC/dt = (q13 * A_0) + (0 * B_0) - (q31 + q34) * C_0 + (q43 * D_0)$$

$$dD/dt = (0 * A_0) + (q24 * B_0) + (q34 * C_0) - (q43 + q42) * D_0$$

At steady state $dA/dt = 0, dB/dt = 0$, etc., and therefore:

$$0 = -(q12 + q13)A + q21B + q31C + 0$$

$$0 = q12A - (q21 + q24)B + 0 + q42D$$

$$0 = q13A + 0 - (q31 + q34)C + q43D$$

$$0 = 0 + q24B + q34C - (q42 + q43)D$$

This set of linear equations can be solved for A, B, C, and D, with the requirement that A+B+C+D=1. We replaced 0 rates with the smallest nonzero rate in the model multiplied by 0.001 to allow transitions between all states. The positive nonzero solution for A, B, C, and D can then be conceived as the expected residence times along some arbitrary bacterial lineage. We used these residence times to estimate a Putative Client Index, PCI, denoting the normalized residence time in state C:

$$PCI(gene) = \frac{C}{(C + D)(A + C)} = \frac{Pr(gene = present \cap hsp90A = absent)}{Pr(gene = present) * Pr(hsp90A = absent)}$$

Notably, if Hsp90A and the gene□s product have no client relationship, the proportion of time spent in state C is expected to be equal to $(C + D) * (A + C)$. Small values of PCI therefore indicate that a gene is observed less frequently than expected without hsp90A. Since no obvious threshold value can be defined, we considered the 20 genes with the lowest PCI values as putative clients (Figure 3 and Table 2; Methods).

To account for variation in rates between BayesTraits runs we repeated this procedure 100 times and defined as putative clients those that were identified as clients in at least 90 of these runs (see Supporting Text). PCI scores shown in Table 2 and Figure 3 are averages across all runs.

7.4.10 *Functional enrichment analysis*

We used a hypergeometric test to assess whether each KEGG functional annotation is overrepresented in the various Hsp90-associated gene classes. As a background set in each case we used the entire set of genes analyzed. Any annotation present in less than 4 copies in the background set was not considered. We accepted enrichments at a 5% FDR.

7.4.11 *E. coli strains and growth assays*

Escherichia coli K-12 strains and plasmids used in this study are listed in Table S2. Cells were grown in tryptone broth (TB; 1% tryptone and 0.5% NaCl) and when necessary supplemented with ampicillin, chloramphenicol and/or kanamycin at final concentrations of 100, 35 and 50 μg/ml, respectively. Overnight cultures, grown at 30¡C, were diluted 1:100 and grown at 34¡C for about 4 h, to an OD600 of 0.45-0.5. All expression constructs for YFP and CFP fusions were constructed as described previously [19,66,67] . Induction levels for protein expression were 1 μM IPTG (pHL24, pHL35, pVS129 and pVS132), 20 μM IPTG (pVS64 and pVS99), 25 μM IPTG (pDK36, pDK90 and pDK91), 50 μM IPTG (pDK19 and pVS18), 0.005% arabinose (pHL13, pVS108 and pVS109) and 0.01% arabinose (pHL52, pHL70, pDK14, pDK29, pDK30 and pDK49). Cells were harvested by centrifugation (4,000 rpm, 5 min), washed once with tethering buffer (10mM potassium phosphate, 0.1 mM EDTA, 1 mM L-methionine, 67 mM sodium chloride, 10 mM sodium lactate, pH 7) and resuspended in 10 mL tethering buffer prior to FRET measurements. TB soft agar plates were prepared by supplementing TB with 0.3% agar (Applichem) and when necessary with 100 g/mL ampicillin

72

and 1 μM IPTG. Equal amounts of cells from different overnight cultures, adjusted depending on their optical density to the equivalent of 2.5 μL of culture with OD600 of 2.0, were inoculated and allowed to spread at indicated temperatures for indicated times. Following incubation, photographs of plates were taken with a Canon EOS 300D (DS6041) camera. Images were analyzed with ImageJ (Wayne Rasband, NIH, http://rsb.info.nih.gov/ij/) to determine the diameter of the rings of spreading colonies. For analysis of motility at different growth stages (indicated by OD600 value), percentages of motile cells were estimated from the microscopy movies of swimming cells. The experiment was performed with the RP437 strain, which is non-motile above 37ºC. Cells were grown overnight in TB medium at 37ºC to completely inhibit their motility. After dilution in fresh TB medium to OD600 0.01, cells were grown at 34ºC for measurements.

7.4.12 *Fluorescence imaging*

For microscopy, cells were taken from the soft-agar plates and applied to a thin agarose pad (1% agarose in tethering buffer). Fluorescence imaging was performed on a Zeiss AxioImager microscope equipped with an ORCA AG CCD camera (Hamamatsu), a 100? NA 1.45 objective, and HE YFP (Excitation BP 500/25; Emission BP 535/30) and HE CFP (Excitation BP 436/25; Emission BP 480/40) filter sets. Each imaging experiment was performed in duplicate on independent cultures. All images were acquired under identical conditions. Images were subsequently analysed using ImageJ software.

7.4.13 *Acceptor photobleaching FRET measurement*

FRET measurements by acceptor photobleaching were performed on a custom-modified Zeiss Axiovert 200 microscope as described before [66]. Briefly, cells expressing YFP and CFP fusions of interest were concentrated about tenfold by centrifugation, resuspended in tethering buffer and applied to a thin agarose pad (1% agarose in tethering buffer). Excitation light from a 75 XBO lamp, attenuated by a ND60 (0.2) neutral-

density filter, passed through a band-pass (BP) 436/20 filter and a 495DCSP dichroic mirror and was reflected on the specimen by a Z440/532 dual-band beamsplitter (transmission 465-500 and 550-640 nm; reflection 425-445 and 532 nm). Bleaching of YFP was accomplished by a 20 sec illumination with a 532 nm diode laser (Rapp OptoElectronic), reflected by the 495DCSP dichroic mirror into the light path. Emission from the field of view, which was narrowed with a diaphragm to the area bleached by the laser, passed through a BP 485/40 filter onto a H7421-40 photon counter (Hamamatsu). For each measurement point, photons were counted over 0.5 s using a counter function of the PCI-6034E board, controlled by a custom-written LabView 7.1 program (both from National Instruments). CFP emission was recorded before and after bleaching of YFP, and FRET was calculated as the CFP signal increase divided by the total signal after bleaching. $\Delta flhC$ strains were used to define direct interactions between HtpG and flagellar and chemotaxis components. In this background expression of endogenous flagellar and chemotaxis genes is inhibited, thus eliminating indirect interactions that may result from concomitant binding of HtpG and tested protein to a third flagellar or chemotaxis protein.

7.5 *Acknowledgments*

Chapter 8

EVOLUTIONARY ASSEMBLY PATTERNS OF PROKARYOTIC GENOMES

A version of this chapter is under review for publication, and is available at:

http://biorxiv.org/content/early/2015/09/27/027649.

Supporting figures and tables can be found in Appendix F.

Chapter 9

THE THESIS UNFORMATTED

This chapter describes the uwthesis class (`uwthesis.cls`, version dated 2011/06/27) in detail and shows how it was used to format the thesis. A working knowledge of Lamport's LaTeX manual[?] is assumed.

9.1 The Control File

The source to this sample thesis is contained in a single file only because ease of distribution was a concern. You should not do this. Your task will be much easier if you break your thesis into several files: a file for the preliminary pages, a file for each chapter, one for the glossary, and one for each appendix. Then use a control file to tie them all together. This way you can edit and format parts of your thesis much more efficiently.

Figure 9.1 shows a control file that might have produced this thesis. It sets the document style, with options and parameters, and formats the various parts of the thesis—but contains no text of its own.

The first section, from the `\documentclass` to the `\begin\{document\}`, defines the document class and options. This thesis has specified two-sided formatting, which is now allowed by the Graduate School. Two sided printing is now actually LaTeX's default. If you want one sided printing you must specify `oneside`. This sample also specified a font size of 11 points. Possible font size options are: `10pt`, `11pt`, and `12pt`. Default is 12 points, which is the preference of the Graduate School. If you choose a smaller size be sure to check with the Graduate School for acceptability. The smaller fonts can produce very small sub and superscripts.

Include most additional formatting packages with `\usepackage`, as describe by Lamport[?].

Figure 9.1: A thesis control file (`thesis.tex`). This file is the input to LaTeX that will produce a thesis. It contains no text, only commands which direct the formatting of the thesis. This is also an example of a 'facing page' caption. It is guaranteed to appear on a lefthand page, facing the figure contents on the right. See the text.

```
% LaTeX thesis control file

\documentclass[11pt,twoside]{uwthesis}

\begin{document}

% preliminary pages
%
\prelimpages
\include{prelim}

% text pages
%
\textpages
\include{chap1}
\include{chap2}
\include{chap3}
\include{chap4}

% bibliography
%
\bibliographystyle{plain}
\bibliography{all}

% appendices
%
\appendix
\include{appxa}
\include{appxb}

\include{vita}
\end{document}
```

The one exception to this rule is the `natbib` package. Include it with the `natbib` document option.

Use the `\includeonly` command to format only a part of your thesis. See Lamport[**?**, sec. 4.4] for usage and limitations.

9.2 The Text Pages

A chapter is a major division of the thesis. Each chapter begins on a new page and has a Table of Contents entry.

9.2.1 Chapters, Sections, Subsections, and Appendices

Within the chapter title use a `\\` control sequence to separate lines in the printed title (recall Figure **??**.). The `\\` does not affect the Table of Contents entry.

Format appendices just like chapters. The control sequence `\appendix` instructs LaTeX to begin using the term 'Appendix' rather than 'Chapter'.

Sections and subsections of a chapter are specified by `\section` and `\subsection`, respectively. In this thesis chapter and section titles are written to the table of contents. Consult Lamport[**?**, pg. 176] to see which subdivisions of the thesis can be written to the table of contents. The `\\` control sequence is not permitted in section and subsection titles.

9.2.2 Footnotes

Footnotes format as described in the LaTeX book. You can also ask for end-of-chapter or end-of-thesis notes. The thesis class will automatically set these up if you ask for the document class option `chapternotes` or `endnotes`.

If selected, chapternotes will print automatically. If you choose endnotes however you must explicitly indicate when to print the notes with the command `\printendnotes`. See the style guide for suitable endnote placement.

9.2.3 Figures and Tables

Standard LaTeX figures and tables, see Lamport[**?**, sec. C.9], normally provide the most convenient means to position the figure. Full page floats and facing captions are exceptions to this rule.

If you want a figure or table to occupy a full page enclose the contents in a `fullpage` environment. See figures 9.2.

Facing page captions are described in the Style Manual[**?**]. They have different meanings depending on whether you are using the one-side or two-side thesis style.

If you are using the two-side style, facing captions are full page captions for full page figures or tables and must face the illustration to which they refer. You must explicitly format both pages. The caption part must appear on an even page (left side) and the figure or table must come on the following odd page (right side). Enclose the float contents for the caption in a `leftfullpage` environment, and enclose the float contents for the figure or table in a `fullpage` environment. Figure 9.1, for example, required a full page so its caption (on a facing caption page) would have been formatted as shown in figure 9.2a. The first page (left side) contains the caption. The second page (right side) could be left blank. A picture or graph might be pasted onto this space.

If instead you are using the one-side style, facing caption pages are still captions for full page figures or tables that appear on the left-hand page (facing the illustration on the right-hand page). However, the page number and binding offset are reversed from their normal positions. Format these captions by enclosing the float contents in a `leftfullpage` environment. Because you are printing on only one side of each sheet, you must manually turn over this caption sheet. You then have the choice of inserting a preprinted illustration or formatting one to print with the thesis. In either case no page number should appear on the illustration page, nor should the page number increment. Enclose your figure's text in an `xtrafullpage` environment, which will cause the page numbers to come out right. You can, of course, leave out the illustration and insert a

```
\begin{figure}[p]% the left side caption
  \begin{leftfullpage}
    \caption{ . . . }
  \end{leftfullpage}
\end{figure}
\begin{figure}[p]% the right side space
  \begin{fullpage}
    . . .
      ( note.. no caption here )
  \end{fullpage}
\end{figure}
```

Figure 9.2: (

a)This text would create a double page figure in the two-side style.

preprinted copy later. Figure 9.2b shows how to format a facing caption page in the one-side style. Note that, in this case, the illustration was also printed.

In the two-side style the `xtrafullpage` environment acts just like the `fullpage` environment. It does not produce a numberless page.

9.2.4 Horizontal Figures and Tables

Figures and tables may be formatted horizontally (a.k.a. landscape) as long as their captions appear horizontal also. LaTeX will format landscape material for you if a couple of conditions are met. You have to have a printer and printer driver that allow rotations and you have to have a couple of add-on LaTeX packages.

Include the `rotating` package

```
\usepackage[figuresright]{rotating}
```

and read the documentation that comes with the package.

Figure 9.4 is an example of how a landscape table might be formatted.

```
\begin{figure}[p]
  \begin{leftfullpage}
    \caption{ . . . }
  \end{leftfullpage}
\end{figure}
\begin{figure}[p]% the right side space
  \begin{xtrafullpage}

    . . .

    ( note.. no caption here )
  \end{xtrafullpage}
\end{figure}
```

Figure 9.3: (

b)[Generating a facing caption page]This text would create a facing caption page with the accompaning figure in the one-side style.

```
\begin{sidewaystable}

   ...

   \caption{ . . . }
\end{sidewaystable}
```

Figure 9.4: This text would create a landscape table with caption.

9.2.5 Figure and Table Captions

Most captions are formatted with the `\caption` macro as described by Lamport[**?**, sec. C.9]. The uwthesis class extends this macro to allow continued figures and tables, and to provide multiple figures and tables with the same number, e.g., 3.1a, 3.1b, etc.

To format the caption for the first part of a figure or table that cannot fit onto a single page use the standard form:

```
\caption[toc]{text}
```

To format the caption for the subsequent parts of the figure or table use this caption:

```
\caption(-){(continued)}
```

It will keep the same number and the text of the caption will be *(continued)*.

To format the caption for the first part of a multi-part figure or table use the format:

```
\caption(a)[toc]{text}
```

The figure or table will be lettered (with 'a') as well as numbered. To format the caption for the subsequent parts of the multi-part figure or table use the format:

```
\caption(x){text}
```

where *x* is b, c, ... The parts will be lettered (with 'b', 'c', ...).

9.3   The Preliminary Pages

These are easy to format only because they are relatively invariant among theses. Therefore the difficulties have already been encountered and overcome by LaTeX and the thesis document classes.

Start with the definitions that describe your thesis. This sample thesis was printed with the parameters:

```
\Title{The Suitability of the \LaTeX\ Text Formatter\\
    for Thesis Preparation by Technical and\\
    Non-technical Degree Candidates}
\Author{Jim Fox}
\Program{UW Information Technology}
\Year{2012}

\Chair{Name of Chairperson}{title}{Chair's department}
\Signature{First committee member}
```

```
\Signature{Next committee member}
\Signature{etc}
```

Use two or more `\Chair` lines if you have co-chairs.

9.3.1   Copyright page

Print the copyright page with `\copyrightpage`.

9.3.2   Title page

Print the title page with `\titlepage`. The title page of this thesis was printed with[1]

```
\titlepage
```

You may change default text on the title page with these macros. You will have to redefine $\degree$text, for instance, if you're writing a Master's thesis instead of a dissertation.[2]

`\Degree{`*degree name*`}`   defaults to "Doctor of Philosophy"

`\School{`*school name*`}`   defaults to "University of Washington"

`\Degreetext{`*degree text*`}`   defaults to "A dissertation submitted ..."

`\textofCommittee{`*committee label*`}`   defaults to "Reading Committee:"

`\textofChair{`*chair label*`}`   defaults to "Chair of the Supervisory Committee:"

These definitions must appear <u>before</u> the `\titlepage` command.

---

[1] Actually, it wasn't. I added a footnote—something you would not do.

[2] If you use these they can be included with the other information before copyrightpage".

9.3.3 *Abstract*

Print the abstract with `\abstract`. It has one argument, which is the text of the abstract. All the names have already been defined. The abstract of this thesis was printed with

```
\abstract{This sample . . . 'real' dissertation.}
```

9.3.4 *Tables of contents*

Use the standard LaTeX commands to format these items.

9.3.5 *Acknowledgments*

Use the `\acknowledgments` macro to format the acknowledgments page. It has one argument, which is the text of the acknowledgment. The acknowledgments of this thesis was printed with

```
\acknowledgments{The author wishes . . . {\it il miglior fabbro}.\par}}
```

BIBLIOGRAPHY

[1] Laura E Dixon, Kirsten Knox, Laszlo Kozma-Bognar, Megan M Southern, Alexandra Pokhilko, and Andrew J Millar. Temporal repression of core circadian genes is mediated through EARLY FLOWERING 3 in Arabidopsis. *Current biology : CB*, 21(2):120–5, January 2011.

[2] Jennifer R Gatchel and Huda Y Zoghbi. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature reviews. Genetics*, 6(10):743–55, October 2005.

[3] Rita Gemayel, Marcelo D Vinces, Matthieu Legendre, and Kevin J Verstrepen. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics*, 44:445–77, January 2010.

[4] Woe-Yeon Kim, Karen A Hicks, and David E Somers. Independent roles for EARLY FLOWERING 3 and ZEITLUPE in the control of circadian timing, hypocotyl length, and flowering time. *Plant physiology*, 139(3):1557–69, 2005.

[5] Elsebeth Kolmos, Eva Herrero, Nora Bujdoso, Andrew J Millar, Réka Tóth, Peter Gyula, Ferenc Nagy, and Seth J Davis. A Reduced-Function Allele Reveals That EARLY FLOWERING3 Repressive Action on the Circadian Clock Is Modulated by Phytochrome Signals in Arabidopsis. *The Plant Cell Online*, 23(9):3230–3246, 2011.

[6] Dmitri A Nusinow, Anne Helfer, Elizabeth E Hamilton, Jasmine J King, Takato Imaizumi, Thomas F Schultz, Eva M Farré, and Steve A Kay. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature*, 475(7356):398–402, July 2011.

[7] Harry T. Orr. Polyglutamine neurodegeneration: Expanded glutamines enhance native functions. *Current Opinion in Genetics and Development*, 22(3):251–255, 2012.

[8] William C Wimsatt. The Units of Selection and the Structure of the Multi-Level Genome. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1980:122–183, January 1980.
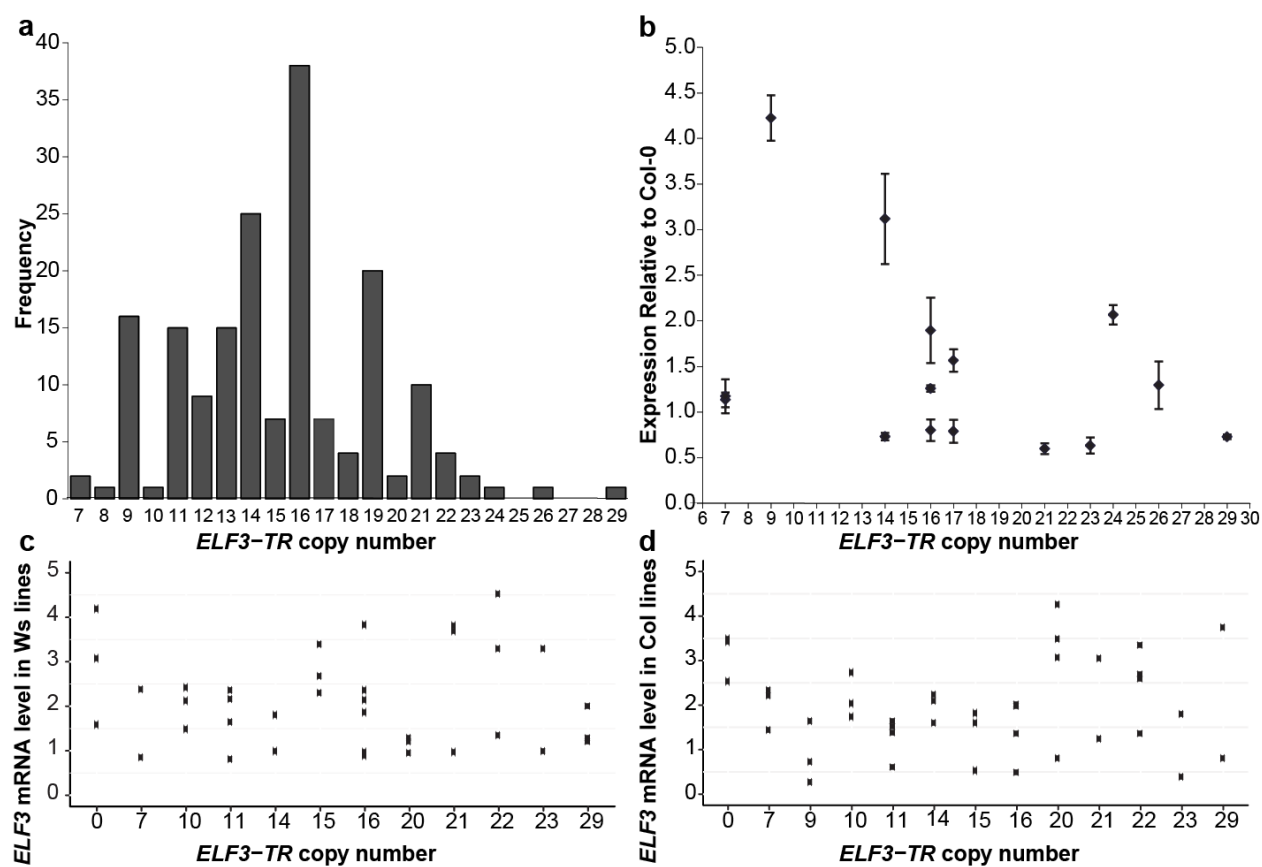
Appendix A

SUPPORTING CHAPTER 2

Figure A.1

Figure A.1: The ELF3-TR variation is not correlated with ELF3 expression. (A) Histogram of ELF3-TR copy number across 181 accessions. TR copy number was determined by Sanger sequencing. (B) ELF3 expression levels in selected natural accessions were measured by quantitative RT-PCR. Expression values are given relative to the Col-0 wild-type reference. Three biological replicates with three technical replicates each were used to obtain expression values. Bars indicate $\pm$ SEM. (C and D) ELF3-TR transgenic lines are expression-matched in both genetic backgrounds. (C) elf3-4, Ws; (D) elf3-200, Col. ELF3 mRNA levels were measured by quantitative PCR (for primers see Table A4) in pooled 10-d-old seedlings that were grown under LD and collected at ZT 20 for each independently generated ELF3-TR transgenic line. ELF3 expression levels are shown relative to either Ws (C) or Col-0 (D) wild-types. Because ELF3 expression levels are known to substantially affect ELF3-dependent phenotypes [4], ELF3 expression is an important variable to consider in our assessment of polyQ tract-length effects. We made efforts to consider only lines within a certain range of ELF3 expression and to test multiple independent lines per ELF3-TR allele (Tables A2□A4), but because of the technical constraints of transgenic plant construction, we cannot entirely exclude the possibility that ELF3 expression partially explains our observations. Although the effects of both ELF3 expression level and ELF3-TR copy number were highly significant, they appear to be largely independent. For example, the ELF3-23Q and ELF3-16Q alleles, which were among the most distinct ELF3-TR alleles in both backgrounds, had very similar ranges of ELF3 expression. In Ws, the alleles ELF3-7Q, ELF3-23Q, and ELF3-10Q phenocopied an elf3 loss-of-function mutant for some phenotypes. Their ELF3 expression levels, however, were very similar to the ELF3-16Q allele, which complemented many ELF3 functions in elf3-4. As observed with individual ELF3-TR alleles, the phenotypic effects of ELF3 expression levels appear to be largely independent of ELF3-TR copy number, which consistently explained a larger portion of phenotypic variation.
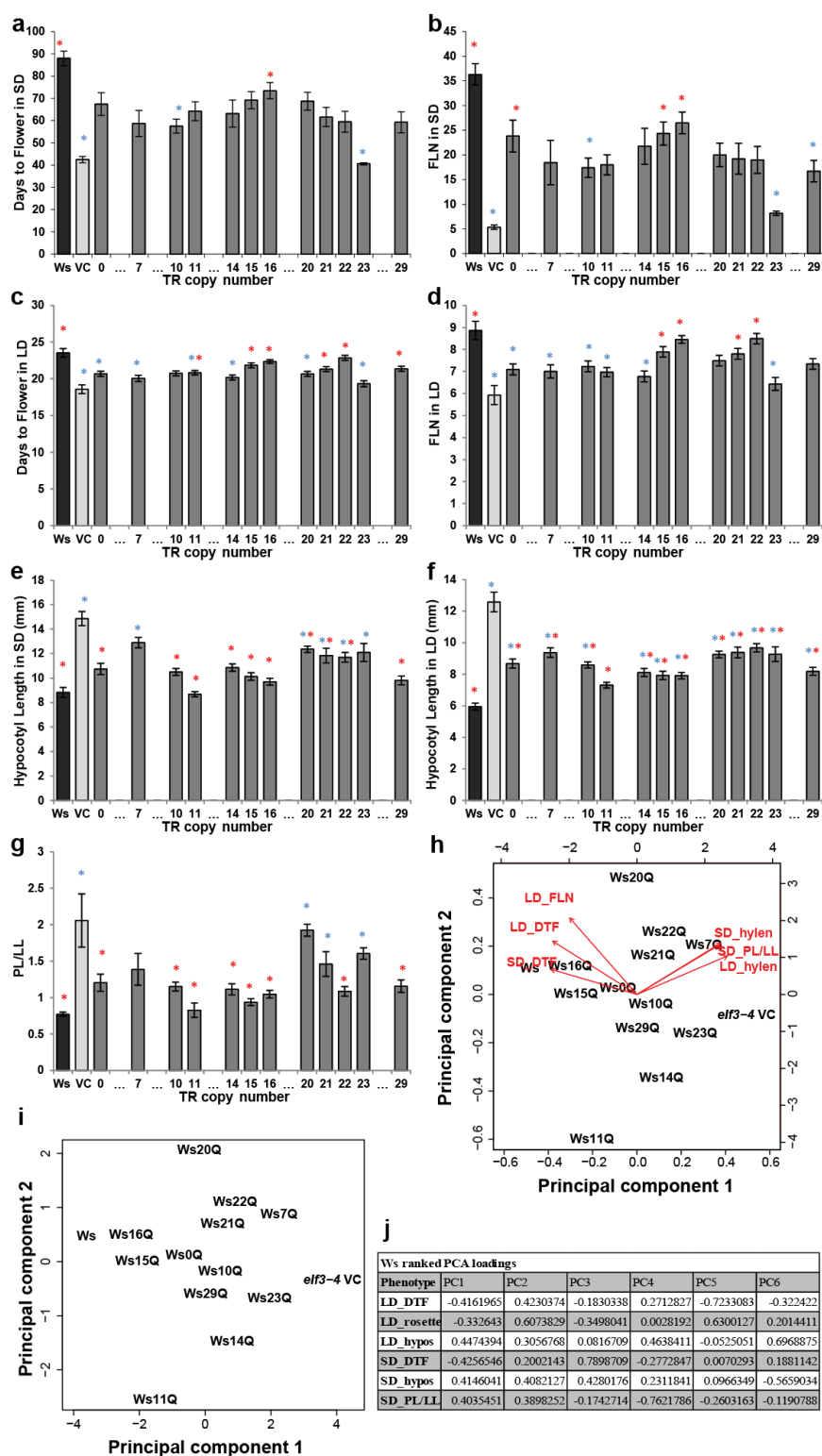

Figure A.2

Figure A.2: ELF3-TR variation has nonlinear phenotypic effects in the elf3-4 background (Ws accession). (A) Days to flower (DTF) under SD (n = 6 plants per line). (B) Final number of rosette leaves (FLN) under SD (n = 6 plants per line). (C) DTF under LD (n = 15 plants per line). (D) FLN under LD (n = 15 plants per line). (E) Hypocotyl length under SD (n = 20-30 seedlings per line). (F) Hypocotyl length under LD (n = 20-30 seedlings per line). (G) PL/LL ratio under SD (n = 6 plants per line). Data are from the same plants as in B. ELF3-TR alleles are indicated with the number of Qs encoded, Ws is wild-type, VC is the elf3-4 vector control. Blue and red asterisks indicate alleles that are significantly different from the wild-type and from the VC, respectively, by Tukey-HSD test ($\alpha = 0.05$). We used this analysis rather than the one presented in Figure 1B to preserve clarity. Bars indicate $\pm$ SEM. These experiments were repeated at least once with similar results. (H and I) PCA of phenotypic data for all ELF3-TR alleles in the elf3-4 background (Ws accession). (H) Biplot of PC1 and PC2, graphically showing the contribution of phenotypes to PCs as red arrows. Note that for the biplot representation, PC1 and PC2 are transformed to the same scale (bottom and left axes), whereas phenotype contributions (in red) are allowed to differ in scale (top and right axes). Phenotypes are hypocotyl length in short and long days (SD hylen and LD hylen), DTF in short and long days (SD DTF and LD DTF), and FLN in long days (LD rosette). Wild-type plants are characterized by late flowering (large SD and LD DTF, many rosette leaves) and short hypocotyls (small SD and LD hylen), relative to elf3 loss-of-function mutants. (I) PC1 and PC2. (J) PCA loadings for Ws background. hylen, hypocotyl length (mm). PCA loadings describe the composition/loading of each principal component . For PC1, flowering-time phenotypes and circadian clock phenotypes have opposite loading signs.
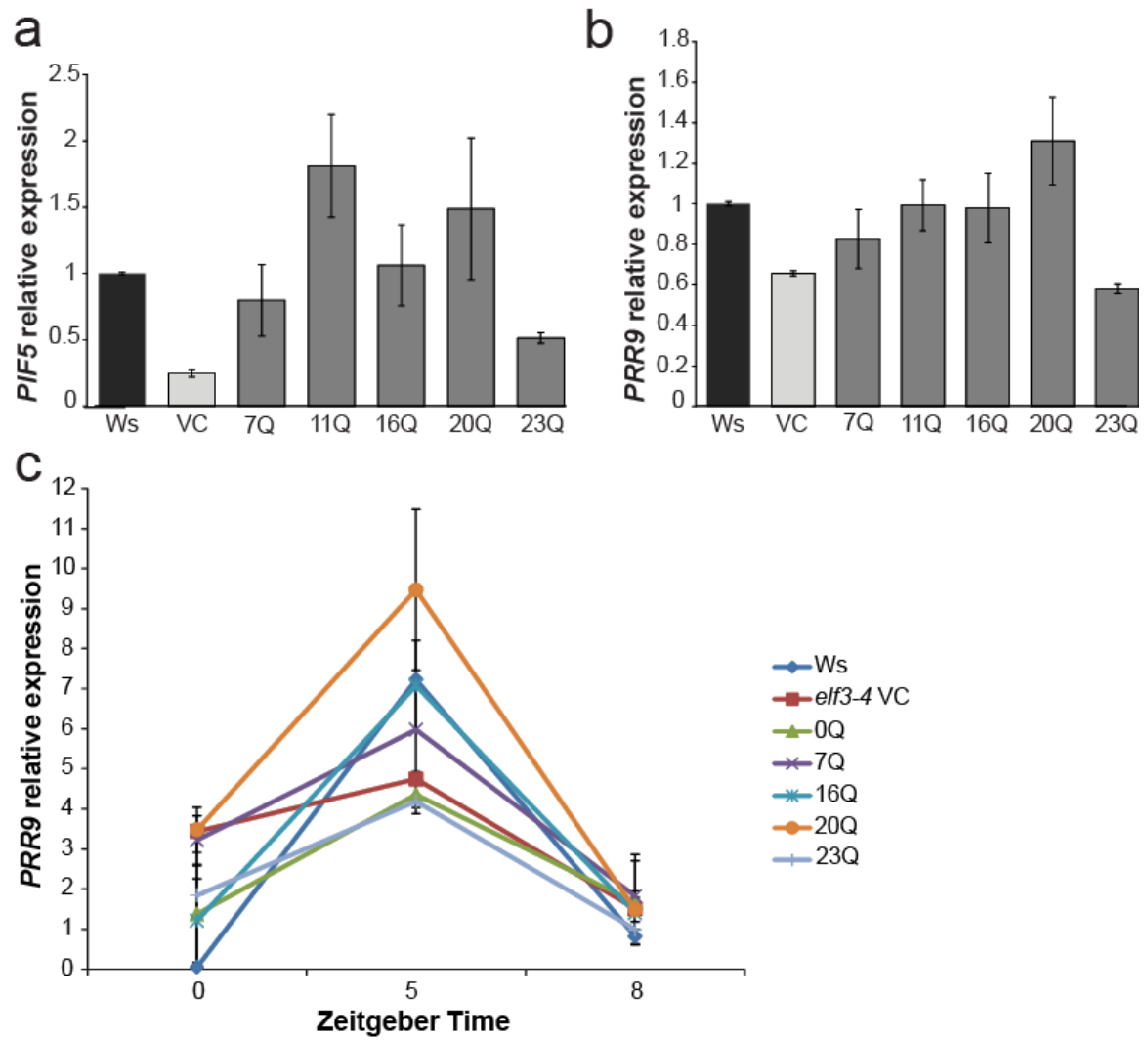
Figure A.3

Figure A.3: Expression levels of the ELF3-regulated genes PIF5 (A) and PRR9 (B and C). Plants were grown under LD and RNA was collected at times showing the largest expression difference between wild-type and elf3-4 mutant ZT8 for PIF5 [6] (A) and ZT5 for PRR9 [5, 1] (B and C). RNA levels were normalized relative to Ws wild-type. (C) Temporal variation in PRR9 expression across ELF3-TR transgenic lines. PRR9 expression levels were measured in 10-d-old plants grown under LD. RNA was collected at times demonstrating the diurnal oscillation of PRR9 expression in wild-type, as determined previously. RNA levels were normalized relative to wild-type (Ws) at ZT8. Gene expression was measured in triplicate for each biological replicate, with multiple independent transgenic lines as biological replicates for each ELF3 allele. Error bars indicate SE of expression across biological replicates. Our expression patterns of PRR9 for wild-type and the elf3-4 mutant are similar to previous observations [5, 1]. ELF3-TR alleles are indicated with the number of Qs encoded, Ws is wild-type, VC is the elf3-4 vector control. Error bars are SEs of means. Data are from multiple independently generated expression-matched (Figure A1C) T3 and T4 lines for each TR copy number allele (Table A2).
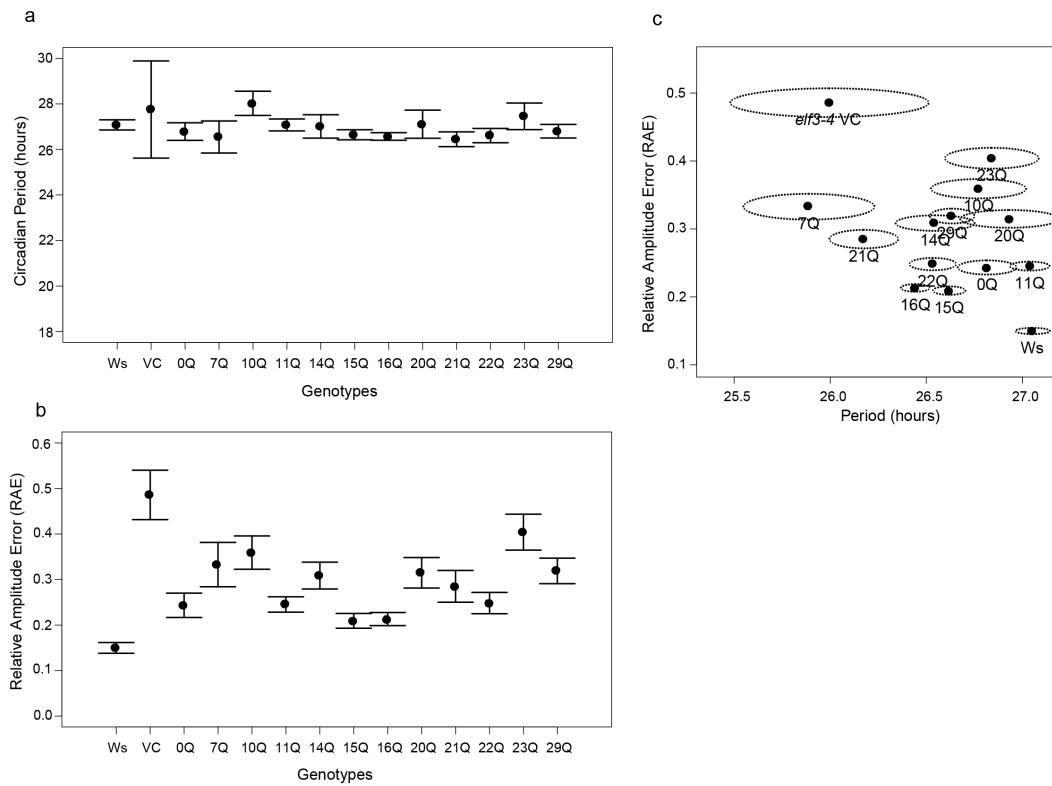
Figure A.4: Circadian parameters estimated for different TR alleles in elf3-4 CCR2::Luc reporter lines. (A) Measured circadian period of CCR::LUC expression oscillation for each ELF3-TR allele. Bars correspond to 99% confidence intervals for this proportion. (B) Measured RAE of CCR::LUC expression oscillation for each ELF3-TR allele. Bars correspond to 99% confidence intervals for this proportion. Plants with RAE <0.4 are considered to have a robust circadian clock. (C) Estimated RAE and circadian period for each ELF3-TR allele. Points are means, dotted ellipses represent SEMs, and genotype labels indicate ELF3-TR copy number. Bioluminescence of the CCR2::LUC reporter present in ELF3-TR transgenic lines was used to measure circadian parameters (period and RAE). Seedlings were entrained in 12-h light:12-h dark cycles for 5 d and released to LL on the sixth day. Note that plants with high RAE have by definition unreliable estimates of circadian period. Number of seedlings for each genotype: Ws, 274; 0Q, 249; 7Q, 122; 10Q, 222; 11Q, 339; 14Q, 214; 15Q, 284; 16Q, 534; 20Q, 161; 21Q, 243; 22Q, 271; 23Q, 196; 29Q, 257; elf3-4 vector control, 102.
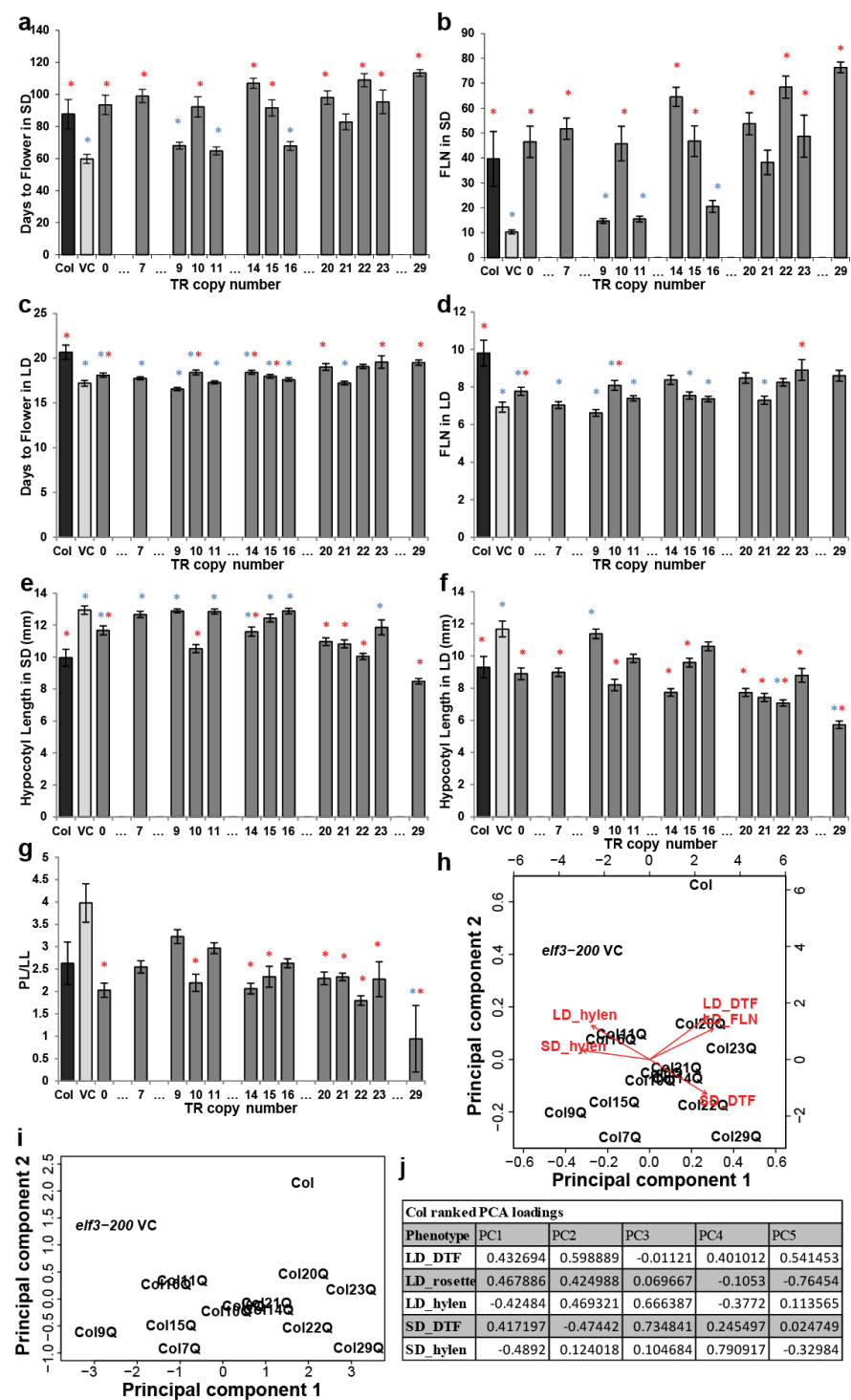
Figure A.5

Figure A.5: ELF3-TR variation has nonlinear phenotypic effects in the elf3-200 background (Col-0 accession). (A) DTF under SD (n = 9 plants/line). (B) FLN under SD (n = 9 plants per line). (C) DTF under LD (n = 15 plants per line). (D) FLN under LD (n = 15 plants per line). (E) Hypocotyl length under SD (n = 20□30 seedlings per line). (F) Hypocotyl length under LD (n = 20□30 seedlings per line). (G) PL/LL ratio under SD (n = 9 plants per line). Data are from the same plants as in B. ELF3-TR alleles are indicated with the number of Qs encoded, Col is wild-type, VC is the elf3-200 vector control (VC). Blue and red asterisks indicate alleles that are significantly different from the wild-type and from the vector control, respectively, by Tukey-HSD test (α = 0.05). Bars indicate ± SEM. These experiments were repeated at least once with similar results. (H and I) PCA of phenotypic data for all ELF3-TR alleles in the elf3-200 (Col accession) background. (H) Biplot of PC1 and PC2, graphically showing the contribution of phenotypes to PCs as red arrows. Note that for the biplot representation, PC1 and PC2 are transformed to the same scale (bottom and left axes), whereas phenotype contributions (in red) are allowed to differ in scale (top and right axes). Phenotypes are hypocotyl length in short and long days (SD hylen and SD hylen), DTF in short and long days (SD DTF and SD DTF), and FLN in long days (SD FLN). Wild-type type plants are characterized by late flowering (large SD and SD DTF, many rosette leaves) and short hypocotyls (small SD and SD hylen), relative to elf3 loss-of-function mutants. (I) PC1 and PC2. Note that PC1's orientation is inverted relative to PCAs including Ws-background plants (A and B: i.e., elf3-200 is to the negative end of the axis, and Col is at the positive end); this does not affect interpretation. In contrast to PCAs including Ws data, PC2 of Col data alone represents the differential response of LD and SD phenotypes to ELF3-polyQ copy number variation. (J) PCA loadings for Col background. hylen = hypocotyl length (mm).

Ws and Col backgrounds ranked PCA

Phenotype	PC1	PC2	PC3	PC4	PC5
LD_DTF	-0.45341	-0.48687	0.21643	-0.06427	0.711617
LD_rosette	-0.44167	-0.5515	0.097755	-0.08292	-0.69595
LD_hylen	0.442085	-0.52945	-0.16168	0.705025	0.032283
SD_DTF	-0.44214	0.09598	-0.87758	0.145134	0.063971
SD_hylen	0.456536	-0.41143	-0.38381	-0.6862	0.064153
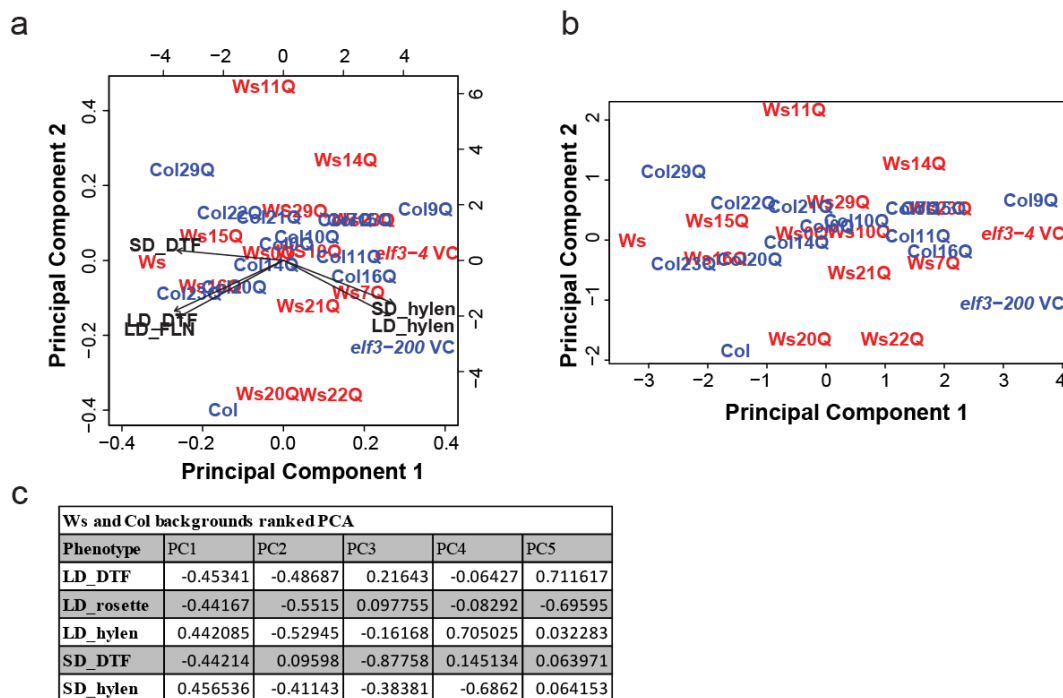
Figure A.6: The phenotypic effects of ELF3-TR copy number variation are strongly background-dependent. PCA of phenotypic data from all ELF3-TR alleles in both elf3-4 (Ws accession) and elf3-200 (Col accession) backgrounds. (A) Biplot of PC1 and PC2, graphically showing the contribution of phenotypes to PCs as black arrows. Note that for the biplot representation, PC1 and PC2 are transformed to the same scale (bottom and left axes), whereas phenotype contributions (in red) are allowed to differ in scale (top and right axes). Phenotypes are hypocotyl length in short and long days (SD hylen and LD hylen), DTF in short and long days (SD DTF and LD DTF), and FLN in long days (LD FLN). Wild-type plants are characterized by late flowering (large SD and LD DTF, many rosette leaves) and short hypocotyls (small SD and LD hylen), relative to elf3 loss-of-function mutants. Text in red represents a given allele in the Ws background (transgenics in elf3-4), and text in blue represents alleles in the Col background (transgenics in elf3-200). (B) PC1 and PC2. (C) PCA loadings for both back- grounds. hylen = hypocotyl length (mm).
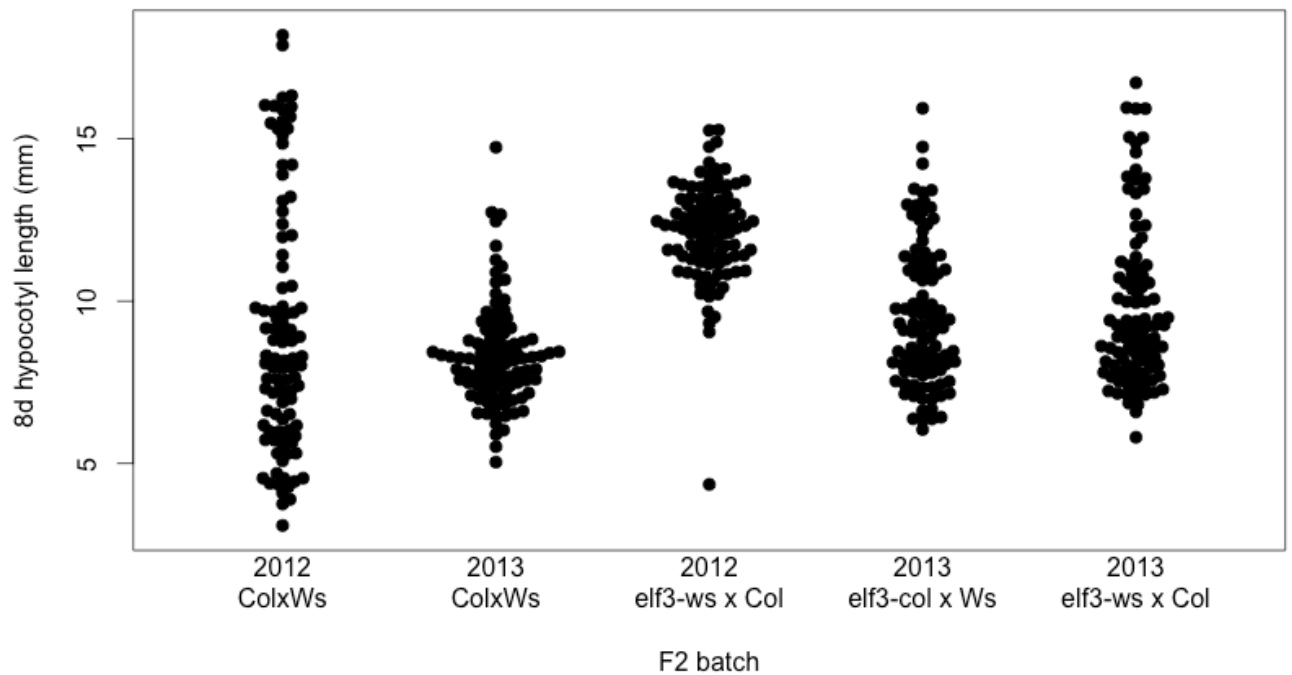
Figure A.7: We measured 8d hypocotyl length under short days in a series of F_2 seedlings generated by various crosses. "2012" batches all involve a single Col individual as a parent (either male or female), which may be a spontaneous mutant. "2013" batches were generated subsequently and represent a variety of Col individuals as parents. "elf3-ws" is the elf3-4 mutant in the Ws background, "elf3-col" is the elf3-200 mutant in the Col background. Hypocotyls ≥ 10 mm are similar to elf3 null mutants, whereas WT hypocotyls are approximately 5-8 mm.

Table A.1: ELF3-TR variation in diverse A. thaliana strains (ecotypes)

Ecotype	ELF3 TR copy number
Ag-0	16
Alc-0	17
Algutsrum	22
An-1	19
Ang-0	12
Ba-1-2	19
Ba-3-3	13
Ba-4-1	17
Bay-0	22
Bg-2	19
Bil-5	16
Bil-7	16
Blh-1	16
Boo-2-1	14
Bor-1	13
Bor-4	13
Br-0	23
Bro-1-6	9
Bs-1	12
Bu-0	19
Bur-0	23
C24	9
Can-0	20
Cen-0	18
CIBC-17	17
CIBC-5	19
Co	13
Co-1	9
Col-0	7
Cvi-0	9
Dem-4	16
Di-0	17
Dra-3-1	16
Drall-1	11
Dralll-1	13
Eden-1	11
Eden-2	11
Edi-0	16
Eds-1	14
Ei-2	15
En-1	11
Es-0	9
Est-0	19
Est-1	19
Fab-2	14
Fab-4	14
Fei-0	19
Ga-0	9
Gd-1	14
Ge-0	18
GOT-22	16
GOT-7	16
Gr-1	21
Gu-0 (=Gue-0)	19
Gul-1-2	11

Gy-0	19		Liarum	16
H55	7		Lillo-1	21
Hi-0	16		Lip-0	15
Hod	12		Lis-1	15
Hov-4-1	16		Lis-2	13
Hovdala-2	16		LL-0	13
HR-10	14		Lm-2	19
HR-5	19		Lom-1-1	11
Hs-0	16		Lov-1	16
Hsm	11		Lov-5	16
In-0	16		LP2-2	26
Is-0	17		LP2-6	21
Jm-0	11		Lu-1	13
Ka-0	19		Lz-0	9
Kas-1	16		Mr-0	24
Kas-2	29		Mrk-0	13
Kavlinge-1	9		MS-0	15
Kent	11		Mt-0	21
Kin-0	19		Mz-0	14
Kni-1	19		N6034	14
Knox-10	16		N6187	19
Knox-18	14		Na-1	22
Koln	12		NC-6	14
Konchezero (N13)	16		Nd-1	16
Kondara	14		NFA-10	18
KZ-1	21		NFA-8	14
Kz-13	14		Nok-3	16
Kz-9	14		Nw-0	14
Lc-0	19		Nyl-2	14
Ler-1	17		Omo-2-1	22

Omo-2-3	21		Shakdara	14
Or-1	9		Sorbo	14
Ost-0	11		Spr-1-2	16
Oy-0	16		Spr-1-6	11
Pa-1	9		SQ-1	14
Per-1	11		SQ-8	15
Petergof	14		St-0	15
Pi-0	12		Stu-1-1	16
Pla-0	9		Stw-0	14
PNA-10	16		Ta-0	21
Pro-0	13		TAMM-2	9
Pu2-23	12		TAMM-27	9
Pu2-7	16		Te-0	10
Pu2-8	11		Tottarp-2	16
Ra-0	12		Ts-1	13
Rd-0	19		Ts-5	13
REN-1	19		Tsu-0	16
Rev-1	19		Tsu-1	16
RMX-A02	16		Tu-0	16
RMX-A180	18		Ull-2-3	15
RRS-10	13		Ull-2-5	11
RRS-7	21		Uod-1	12
Rsch-4	11		Uod-7	21
Rubezhnoe-1	14		Van-0	13
San-2	20		Var-2-1	9
Sanna-2	16		Var-2-6	9
Santa Clara	8		Vimmerby	16
Sap-0	16		Wa-1	16
Se-0	14		Wei-0	12
Sf-1	9		Wil-2	14

Ws-0	21
Ws-2	16
Wt-5	14
Yo-0	16
Zdr-1	13
Zdr-6	17

Table A.2: Expression of ELF3 in transgenic lines (Ws background)

TR copy number	Line	Relative expression	CV(expression)
Vector	V1-1	0.45	0.04
0	0R1-3	4.18	0.01
0	0R4-3	3.08	0.02
0	0R5-3	1.58	0.01
7	7R3-1	2.38	0.02
7	7R5-2	0.86	0.01
10	10R1-1	2.41	0.01
10	10R2-2	1.49	0.01
10	10R4-2	2.13	0.01
11	11R1-3	2.17	0.03
11	11R2-3	1.65	0.01
11	11R8-1	0.81	0.02
11	11R9-1	2.35	0.03
14	14R1-1	0.99	0.01
14	14R4-2	1.81	0.01
15	15R1-2	2.68	0.01
15	15R2-1	2.31	0.02
15	15R3-1	3.40	0.01
16	16R1-2	1.86	0.01
16	16R4-2	2.14	0.01
16	16R6-2	3.82	0.02
16	16R7-2	2.35	0.02
16	16R8-1	0.97	0.02
16	16R10-3	0.89	0.01
20	20R1-2	1.20	0.02
20	20R2-1	0.95	0.02
20	20R3-2	1.28	0.02
21	21R2-2	0.97	0.01
21	21R3-2	3.81	0.01
21	21R5-3	3.69	0.01

22	22R3-3	4.53	0.01
22	22R6-1	1.35	0.01
22	22R8-1	3.29	0.01
23	23R2-1	3.30	0.02
23	23R5-1	0.98	0.00
29	29R3-2	1.29	0.00
29	29R4-1	2.01	0.02
29	29R5-2	1.21	0.02

Table A.3: Expression of ELF3 in transgenic lines (Col background)

TR copy number	Line	Relative expression	CV(expression)
Vector	V1-1	0.05	0.00
Vector	V1-1	0.05	0.00
0	0R1-1	3.48	0.12
0	0R3-1	2.54	0.32
0	0R5-1	3.42	0.42
7	7R3-1	2.33	0.20
7	7R4-3	2.21	0.33
7	7R5-1	1.44	0.03
9	9R2-1	0.27	0.02
9	9R4-1	0.73	0.10
9	9R5-1	1.63	0.21
10	10R1-2	1.74	0.01
10	10R3-3	2.74	0.09
10	10R7-3	2.04	0.25
11	11R3-1	1.38	0.09
11	11R5-4	1.52	0.16
11	11R6-1	1.63	0.21
11	11R7-1	0.61	0.03
14	14R3-4	2.10	0.13
14	14R5-4	2.24	0.00
14	14R10-3	1.59	0.49

15	15R2-2	1.59	0.27
15	15R4-2	1.81	0.08
15	15R8-1	0.53	0.34
16	16R1-1	0.49	0.00
16	16R2-4	2.01	0.12
16	16R3-2	1.37	0.36
16	16R4-2	1.97	0.17
20	20R1-2	4.27	0.39
20	20R2-3	0.81	0.02
20	20R3-3	3.07	0.17
20	20R5-3	3.49	0.57
21	21R1-2	3.05	0.13
21	21R3-2	1.25	0.04
22	22R3-3	3.34	0.26
22	22R4-2	1.37	0.35
22	22R6-1	2.69	0.10
22	22R7-2	2.59	0.54
23	23R2-4	0.39	0.12
23	23R4-3	1.79	0.03
29	29R4-2	0.81	0.21
29	29R11-1	3.75	0.60

Table A.4: Primer sequences

Primer Name/target	Sequence
Nar1 Sense	GAGATCTGATAATGAACCGGCGCCACAGCAACAGCAACAG
Nar1 Antisense	TGTTGCTGTTGCTGTGGCGCCGGTTCATTATCAGATCTC
Nco1 Sense	CTCAATATCACCCCGCCATGGGATTCCCACC
Nco1 Antisense	GGTGGGAATCCCATGGCGGGGTGATATTGAG
PolyQ Sense	CCCTTCCCATGGGATTCCCACCTCCTGGTAAT
PolyQ Antisense	TTTTGGGGCGCCGGTTCATTATCAGATCTCTG
0Q Sense	ACCATAATGAACCCATATTGTTCAAGCCC
	CAATGAGCAAATGAACCAGTTTGGA

oQ Antisense	TCCAAACTGGTTCATTTGCTCATTGGG
	GCTTGAACAATATGGGTTCATTATGGT
ELF3 Construct S	CAATAATGGTTTCTGACGTA
ELF3 Construct A	ACCAATGGTACTCAAAATAGTTTGGTCATACGG

b: Real-Time PCR primers.

Target Gene/orientation	Primer
ELF3F	GACATTGATAATGATCGTGAATACAG
ELF3R	CTAATATACCCACAACATCATCGG
PIF5F	AGTCGGACCGAGTCATT
PIF5R	TCTTTGTTGTTCCTTCCATAGC
PRR9F	ATAAGCTGATGGAGAATGGC
PRR9R	TCCAAGCTCAGGACCAACA
UBC21F	GACCAAGATATTCCATCCTA
UBC21R	GTTAAGAGGACTGTCCG

c: TAIL-PCR primers.

Name	Sequence
LB-0a	GCTGGACTTCAGCCTGCCGGTGCCGCCC
LB-1a	ACGATGGACTCCAGTCCGGCCCCCGTCAC
	CGAAATCTGATGACCCCTAGAGTC
LB-2a	CGCGCGCGGTGTCATCTACTATGTTACTAGATC

Appendix B

SUPPORTING CHAPTER 3

Appendix C

SUPPORTING CHAPTER 4

Appendix D

SUPPORTING CHAPTER 5

Appendix E

SUPPORTING CHAPTER 6

Appendix F

SUPPORTING CHAPTER 7

Appendix G

WHERE TO FIND THE FILES

The uwthesis class file, `uwthesis.cls`, contains the parameter settings.